UNIVERSIDADE FEDERAL DE SANTA CATARINA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

José Gonçalo dos Santos

Uso de Conjuntos Difusos e Lógica Difusa para Cálculo de Atração e Repulsão: Uma Aplicação em *Market Basket Analysis*

Tese apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Santa Catarina como requisito parcial para obtenção do título de Doutor em Ciência da Computação

Profª. Silvia Modesto Nassar, Drª - Orientadora

Florianópolis, dezembro de 2004

Uso de Conjuntos Difusos e Lógica Difusa para Cálculo de Atração e Repulsão: Uma Aplicação em *Market Basket Analysis*

José Gonçalo dos Santos

Esta Tese foi julgada adequada para a obtenção do título de **Doutor em Ciência da Computação** na Área de Concentração Sistemas do Conhecimento e aprovada em sua forma final pelo **Programa de Pós-Graduação em Ciência da Computação** da Universidade Federal de Santa Catarina.

Florianópolis, de de 2004

_____

Raul Sidnei Wazlawick, Dr.
Coordenador do Programa

Banca Examinadora:

_____	_____
Prof ª Silvia Modesto Nassar, Dra. Orientadora	Profº Paulo Martins Engel, Dr.
_____	_____
Prof º Mauro Roisemberg, Dr.	Profº Rogério Cid Bastos, Dr.
_____	_____
Prof º Mário Antonio Ribeiro Dantas, Dr.	Prof º Rui Seara, Dr.

_____

Prof º Paulo A. Bracarense Costa, Dr.

Agradecimentos

Quando se desenvolve um trabalho desta magnitude é preciso contar com a ajuda e compreensão de muitas pessoas ou grupos de pessoas. Por isso, uma página de agradecimento não seria suficiente para agradecer a todos que colaboraram para o término desta missão. Porém, é importante fazer alguns destaques que são:

- à minha orientadora, Silvia Modesto Nassar, pela paciência e tolerância nesta difícil tarefa que foi me orientar e me apoiar nos momentos mais complicados do processo;
- ao PPGCC que acreditou em minha capacidade e me deu esta grande oportunidade;
- à secretaria do PPGCC pela atenção e competência;
- à minha família pela compreensão da importância do meu trabalho;
- aos professores do PPGCC pelos conhecimentos que me repassaram;
- aos membros da banca pelas contribuições em vários aspectos deste trabalho.

Resumo

Recentes avanços na forma de aquisição de dados têm mostrado uma revolução de aumento de capacidade tecnológica de armazenamento destes. Notificaçõs de servidores *web*, dados de transações de clientes, compras com cartão de crédito, uso de cartão fidelidade, entre outros, produzem *terabytes* de dados, diariamente, que são úteis como dados históricos, mas não tão úteis quanto poderiam ser se fossem efetivamente processados de forma que pudessem fornecer padrões e tendências. Esses padrões e as tendências são conhecimentos extraídos (descobertos) desses dados. A Descoberta de Conhecimento em Base de Dados (DCBD) é um campo interdisciplinar de pesquisa que mescla conceitos de estatística, de inteligência artificial e de banco de dados. O seu estudo é motivado pelo crescimento da complexidade, e da quantidade de dados oriundos de todas as esferas do domínio humano e da necessidade de extrair informações úteis dos dados coletados. A descoberta de regras de associação é uma área da DCBD que tem por objetivo encontrar conjuntos de itens freqüentes em transações de uma base de dados e inferir regras capazes de mostrar como um conjunto de itens sofre influência na presença de outros conjuntos de itens. O uso de regras de associação no processo de DCBD tem sido utilizado por diversos pesquisadores. Contudo, os modelos para descoberta de regras de associação trabalham com medidas numéricas. No cálculo das medidas de atração/repulsão, esses métodos utilizam uma base de dados, considerando a ocorrência ou não do evento. Trabalhando dessa forma com uma matriz denominada de matriz de co-ocorrência, que contém valores binários onde 0 (zero) representa a não ocorrência e 1 (um), a ocorrência do evento. Porém, essa matriz utilizada para o cálculo de atração/repulsão entre produtos, com valores binários, despreza a intensidade da associação dos eventos e a quantidade de produtos comprados. Dessa forma, a matriz de co-ocorrência utilizada para o cálculo das medidas de associação não reconhece a imprecisão da ocorrência ou não ocorrência conjunta dos eventos. Para o tratamento da imprecisão podem ser utilizadas a teoria dos conjuntos difusos e da lógica difusa. A modelagem da imprecisão utilizando a abordagem difusa parece ser adequada para tratar o problema da imprecisão presente, não considerada na matriz de co-ocorrência. Assim, esta pesquisa teve por objetivo verificar a adequação da abordagem difusa para modelar a imprecisão contida na matriz de co-ocorrência

utilizada no cálculo da medida atração/repulsão, para propor um modelo difuso para o cálculo de atração/repulsão.

Para a modelagem do método proposto foi necessária a identificação dos métodos mais usados em MBA e a identificação dos modelos de regras usados na lógica difusa; a construção de conjuntos difusos para representar termos lingüísticos usados para as variáveis de entrada e a adequação dos limites dos intervalos das funções de pertinência.

Foram avaliadas várias combinações de funções de pertinência em conjunto com os principais modelos de regras, usando várias amostras de associações entre produtos oriundas de base de dados de três segmentos comerciais. A partir daí, foi proposto um método que mapeia entradas numéricas de freqüências para termos lingüísticos e que possibilita como saída a classificação de associação. Podendo ser de atração ou repulsão, com grau de associação baixa, moderada ou alta

O método mostrou bons resultados e pode ser aplicado na área comercial para análise de dados históricos de vendas. Além disso, pode ser usado nos pontos de vendas para auxiliar o atendente a oferecer um novo produto a determinados clientes, baseado na sua compra atual, porque a resposta do sistema pode ser dada em linguagem natural, o que torna acessível a qualquer usuário do sistema. Pode-se também usar o método para fazer consultas usando linguagem natural.

Palavras-chave: Descoberta de Conhecimento em Base de Dados, Mineração de Dados, *Market Basket Analysis*, Lógica Difusa

Abstract

Recent advances in the data acquisition manner have been revolutionizing the technological capacity of their storage. Web servers logs, data of clients transactions, purchase with credit card, fidelity card use, among others, produce terabytes of data, daily, which are useful as historical data, but not so useful as if they could be if they were indeed processed so that they could supply standards and tendencies. These standards and tendencies are knowledge extracted (discovered) from this data. The Knowledge discovery in database (KDD) is an interdisciplinary field of research that mixture concepts of statistics, artificial intelligence and database. Its study is motivated by the growth of the complexity, and of the quantity of data derived from all the spheres of the human domain and of the need to extract useful information from the collected data. The discovery of association rules is a KDD's area, which has for goal find sets of frequent items in transactions of a database and to infer rules able to show as a set of items suffers influence in the presence of other joint of items. The association Rules use in KDD's process has been used by several researchers. However, the models for discovery of association rules work with numeric measures. In the calculation of attraction/repulsion measures, these methods use a database, built considering the occurrence or non-occurrence of the event. Working, thus, with a matrix denominated matrix of co-occurrence, which contains binary values, where 0 (zero) represents the non-occurrence and 1 (one), the event's occurrence. However this matrix used to calculate attraction/repulsion between products with binary values not consider the intensity of the association of the events and the quantity of products bought. Thus, co-occurrence's matrix used to calculate the association measures does not recognize the imprecision of the occurrence or non-joint occurrence of the events. For the treatment of the imprecision, the theory of the fuzzy sets and the fuzzy logic can be used. The modeling of the imprecision using the fuzzy approach seems to be adequate to treat the problem of the present imprecision, not considered in the co-occurrence's matrix. This way, this research had as goal to verify the adaptation of the fuzzy approach to model the imprecision contained in the co-occurrence matrix used in the calculation of the

measure attraction/repulsion, to propose a fuzzy model for attraction/repulsion calculation.

For modeling of the proposed method it was necessary the identification of the most used methods in MBA and the identification of the rule models used in the fuzzy logic; the construction of fuzzy sets to represent linguistic terms used to the entrance variables and the adaptation of the limits of the intervals of the pertinence functions.

Several membership functions, together with the main rule models, were tested, using several association samples between arising products of database of three commercial segments. From there onwards, it was proposed a method that carries numeric values for linguistic terms and enables the answer supply in natural language to the user.

The method presented good results and can be applied in the commercial area for sales historical data analysis. Moreover, it can be used in the sales points to assist the attendant to offer a new product to certain clients, based on its current purchase, because the answer of the system can be given on natural language, what it turns accessible to any user of the system. The method can also be used to provide consulting using natural language.

Key-Words: Knowledge Discovery Database, Data Mining, Market Basket Analysis, Fuzzy Logic

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas

DCBD - Descoberta de Conhecimento em Base de Dados

DW – Data WareHouse

FRA – Freqüência Relativa de A (Antecedente da regra “Se A então B”)

FRB – Freqüência Relativa de B (Conseqüente da regra “Se A então B”)

FREAB – Freqüência Relativa Esperada de A e B

FROAB – Freqüência Relativa Obtida de A e B

IA – Inteligência Artificial

MBA – *Market Basket Analysis*

MDCAR – Método Difuso para Análise de Histórico de Vendas

CAPÍTULO 1 - INTRODUÇÃO

Recentes avanços na forma de aquisição de dados têm revolucionado a capacidade tecnológica de armazenamento destes. *Logs* de servidores *web*, dados de transações de clientes, compra com cartão de crédito, uso de cartão fidelidade, entre outros, produzem *terabytes* de dados, diariamente, que são úteis como dados históricos, mas não tão úteis quanto poderiam ser se fossem efetivamente processados de forma que pudessem fornecer padrões e tendências (BEKER & VIKTOR, 2004 p. 1). Esses padrões e tendências são conhecimentos extraídos (descobertos) desses dados.

A Descoberta de Conhecimento em Base de Dados (DCBD) é um campo interdisciplinar de pesquisa que mescla conceitos de estatística, de inteligência artificial e de banco de dados (HAN & KAMBER, 2001 p. xix). O seu estudo é motivado pelo crescimento da complexidade, e da quantidade de dados oriundos de todas as esferas do domínio humano e da necessidade de extrair informações úteis dos dados coletados (VELOSO et. al 2001, p. 81).

A descoberta de regras de associação é uma área da DCBD que tem por objetivo encontrar conjuntos de itens freqüentes em transações de uma base de dados e inferir regras capazes de mostrar como um conjunto de itens influencia a presença de outros conjuntos de itens (VELOSO et. al, 2001, p. 81). Ressalta-se que a associação pode ser positiva ou negativa, isto é, uma relação com presença-presença ou presença-ausência de itens. A relação presença-presença é chamada de atração e a presença-ausência é chamada de repulsão entre itens.

O uso de regras de associação no processo de DCBD foi introduzido inicialmente por AGRAWAL et al (1993). A partir daí, muitos trabalhos nessa área têm sido desenvolvidos, dos quais podem ser citados AGGARWAL & YU (1998), AGRAWAL et al (1994), BRIN (1997), HAN & FU (1995), LAKSHMANAN et al (1998), PARK et al (1995), RASTOGI & SHIM (1998) e SRIKANT et al (1998). Todos os modelos usados nestes trabalhos foram baseados na medida de suporte e de confiança e não tratam da repulsão entre itens, ou seja, da associação negativa. Todavia, em

SAVASARE et al (1998), AUSLENDER (2000), HAN & KAMBER (2001), GROTH (2000) e BERRY & LINOFF (1997) podem ser encontrados modelos que tratam deste tipo de associação.

Os modelos para descoberta de regras de associação trabalham com medidas numéricas. No cálculo das medidas de atração/repulsão, esses modelos utilizam uma base de dados construída, considerando a ocorrência ou não do evento. Trabalhando dessa forma com uma matriz denominada de matriz de co-ocorrência, que contém valores binários [0;1], onde 0 (zero) representa a não ocorrência e 1 (um), a ocorrência do evento. Porém, essa matriz utilizada para o cálculo de atração/repulsão entre produtos com valores binários despreza a intensidade da associação dos eventos. Por exemplo, ao analisar dados de vendas de produtos em supermercados, a quantidade de produtos comprada (ARIA et al., 2002) não seria considerada na matriz de co-ocorrência. Isto é, se um produto A teve 6 unidades compradas e um outro produto B teve 3 unidades compradas por um mesmo indivíduo, então a força de atração entre eles seria de 2 para 1 e não de 1 para 1, como estaria representado na matriz de co-ocorrência.

Dessa forma, a matriz de co-ocorrência utilizada para o cálculo das medidas de associação em sua forma binária despreza a força de atração ou repulsão entre eventos. Esta força poderia ser forte, moderada ou fraca, caracterizando-se como uma variável lingüística sob a presença da imprecisão quanto à ocorrência ou não ocorrência conjunta de eventos.

Para o tratamento da imprecisão podem ser utilizadas a teoria dos conjuntos difusos e a lógica difusa. Exemplos da utilização da abordagem difusa são: CURY (2003), para classificação desempenho de transporte urbano; DRAESEKE (1999), para tratamento de dados econômicos; DUALIBE (2001), para processamento de sinais; GUIMARÃES (2000), para dosimetria de pena para crimes previstos nos artigos de 121 a 359 do Código Penal Brasileiro; MATTHEWS (2002), para desenvolvimento de sistemas especialistas; ORTEGA (2001), para problemas de biomedicina na área de epidemiologia; RESSOM et al. (2003), para análise de *cluster*; RIBEIRO & MOREIRA

(2003), para consulta a base de dados comerciais; ROYES (2003), para análise de políticas; SCREMIN (2003), para seleção de componentes principais em análise estatística; STURM et al. (2003), para classificar imagens geográficas e WOOLF & WANG (2000), para análise de dados de carga genética. Nestes trabalhos, a abordagem difusa foi utilizada com o objetivo de transformar valores numéricos (*numérico*) em descritores (difuso) qualitativos, para que possam ser analisados através de regras.

Com esta visão, cabe levantar a seguinte questão de pesquisa que norteia este trabalho: No processo de descoberta de conhecimento em bases de dados é adequado utilizar a abordagem difusa para modelar a imprecisão na matriz de co-ocorrência?

Dessa forma, o problema de pesquisa nesta tese é verificar a adequação da abordagem difusa para modelar a imprecisão contida na matriz de co-ocorrência utilizada no cálculo da medida atração/repulsão.

1.1– Objetivos

Objetivo Geral:

Desenvolver um método para tratar a imprecisão na matriz de co-ocorrência utilizada no cálculo de atração/repulsão entre itens.

Objetivos Específicos:

- Investigar a modelagem difusa de valores numéricos;
- Verificar qual é o processo mais adequado de inferência difusa;
- Decidir, através de testes com vários modelos de regras, qual o modelo de regras é o mais adequado;
- Validar o método proposto.

1.3 – Contribuição da Pesquisa

O problema de regras de associação, atração/repulsão, vem sendo trabalhado ao longo dos anos por pesquisadores como SAVASARE et al (1998) e MENDES (2002). Autores como BERRY & LINOFF (1997, p. 107 - 109) e GROTH (2000, p. 87 - 88) descrevem métodos que são usados para tratar deste problema. O grande inconveniente desses métodos é que eles mapeiam entradas numéricas para saídas numéricas que requerem certa especialidade por parte do usuário para descobrir o significado dos números apresentados.

Para mapear entradas numéricas para termos qualitativos de forma que possam ser analisados através de regras, pesquisadores como WOOLF & WANG (2000) e RIBEIRO & MOREIRA (2003) empregaram com sucesso a teoria dos conjuntos difusos.

Nesse contexto, como contribuição principal, a presente pesquisa propõe uma nova abordagem para o cálculo de atração/repulsão, utilizando conjuntos difusos para mapear valores numéricos para termos qualitativos. Para isto, a pesquisa investiga e explora os diversos modelos difusos de inferência, diversas funções de pertinência e também várias combinações dessas funções. Outra contribuição importante é o estudo comparativo feito entre os modelos de *Mamdani*, *Larsen*, *Takagi-Sugeno* e *Tsumamoto*, mostrando as dificuldades, as vantagens e as desvantagens no uso de cada um deles e, também, as sugestões de adaptação dos modelos ao problema em questão são apresentadas.

1.4 – Estrutura do trabalho

Esta tese está dividida em cinco capítulos. No primeiro capítulo, consta a introdução, na qual são apresentados a motivação, os objetivos propostos e a contribuição da pesquisa.

No segundo capítulo é apresentada a revisão da literatura, considerada necessária para o desenvolvimento desta pesquisa. Inicia-se por uma abordagem de Descoberta de Conhecimento em Base de Dados e suas etapas, dando ênfase à etapa de *Mineração de dados* e suas técnicas. São apresentados também fundamentos de *Market Basket Analysis* (MBA) e, por fim, é apresentada a teoria de lógica difusa.

No capítulo três é apresentado, descrito passo a passo e discutido o método proposto nesta pesquisa.

O capítulo quatro é destinado à apresentação e à discussão dos resultados obtidos com o método na etapa de testes.

As considerações finais, com base no desenvolvimento do método, bem como as propostas de trabalhos futuros são apresentadas no capítulo cinco.

Capítulo 2 – Base Conceitual

Neste capítulo são apresentados os fundamentos teóricos que deram suporte à pesquisa, começando-se pela Descoberta de Conhecimento em Base de Dados (DCDB), Mineração de Dados, *Market Baket Analysis* e, finalmente, Lógica Difusa (*Fuzzy Logic*).

2.1 – Descoberta de Conhecimento em Base de Dados (DCBD)

Ao inciar o assunto de descoberta de conhecimento em base de dados (DCBD), cabe a seguinte pergunta: DCBD tem o mesmo significado de *Mineração de dados* (DM)? Em realidade, não há um consenso em relação a isso, pois, para alguns autores, os dois termos têm o mesmo significado. Segundo GROTH (2000, p. 3-4), esses temas estão abertos à debate, e a definição de cada um pode variar, dependendo do autor escolhido para leitura.

Para AMARAL (2001, p. 3), a busca por padrões úteis, em base de dados tem recebido diversos nomes; como descoberta de conhecimento em base de dados, mineração de dados, descoberta de informação, arqueologia dos dados ou processo de padronização de dados. Enquanto que o termo mineração de dados é usado pelos estatísticos e analistas de dados, os pesquisadores de Inteligência Artificial (IA) utilizam o termo DCBD.

Nesta pesquisa é feita uma distinção em relação aos dois termos. DCBD é tratado como o processo completo de descoberta de conhecimento em base de dados e mineração de dados é tratada como uma etapa desse processo e é apresentada na Seção 2.3. Nesta seção são apresentadas a definição, as etapas e as aplicações do DCBD.

2.1.1 – Definição

DCBD (Descoberta de Conhecimento em Base de Dados), de acordo com AMARAL (2001, p.13), “é a descoberta de novos conhecimentos, que podem ser padrões, tendências, associações, probabilidades ou fatos que não são óbvios ou de fácil identificação”. Já GROTH (2000, p. 3) define mineração de dados como sendo a busca por tendências e padrões em base de dados. As definições de GROTH (2000, p. 3) e AMARAL (2001, p. 13) levam a crer que ambos os termos têm o mesmo significado. Mas, para KLÖSGEN & ZYTKOW (2002, p. 2), o termo DCBD se refere a todo o processo de descoberta de conhecimento em dados, enquanto mineração de dados é vista como um passo central desse processo, que aplica algoritmos para extrair e verificar hipóteses.

DCBD é um problema multidisciplinar, que envolve inteligência artificial, estatística, visualização, banco de dados e aprendizagem de máquina, mas segundo KLÖSGEN & ZYTKOW (2002, p. 22), a ciência, a filosofia da ciência e a lógica têm um papel muito importante para a origem do DCBD, porque são responsáveis pelos conceitos básicos de dados, do conhecimento, da linguagem formal e do raciocínio lógico.

A descoberta de conhecimento em base de dados envolve diversas etapas que estão descritas a seguir e ilustradas na figura 2.1.

2.1.2 – Etapas do DCBD

O processo de DCBD começa com a definição do objetivo do problema em questão. Alguns autores definem esta etapa como pertencente à fase de mineração de dados, como é o caso de CARVALHO (2001, p. 11). Já GROTH (2000, p. 46) diz que o processo de DCBD começa com a preparação dos dados. Para AMARAL (2001, p. 15), a definição do objetivo é a definição do conhecimento que o usuário deseja obter sobre os dados. É nessa etapa que é definido o tipo de padrão que se deseja descobrir na base

de dados. Nesta pesquisa, o processo de DCBD inicia-se com a definição dos objetivos, conforme ilustra a figura 2.1.

A segunda etapa do processo é a aquisição de dados, que pode ser feita com o auxílio de um *Data Warehouse* (DW). A utilização de Data Warehouse é defendida por alguns autores (HAN & KAMBER, 2001, p. 39-99) e negada por outros (GROTH,2000, p. 48).

A Figura 2.1 mostra as demais etapas do DCBD, começando pela definição dos objetivos e seguindo pela seleção dos dados, pré-processamento, transformação, mineração e interpretação. Cada etapa descrita a seguir tem um papel importante no processo de DCBD. Uma observação deve ser feita: as etapas de preparação dos dados, segundo AMARAL (2001, p. 17), consomem 70% do tempo destinado ao processo de descoberta de conhecimento.

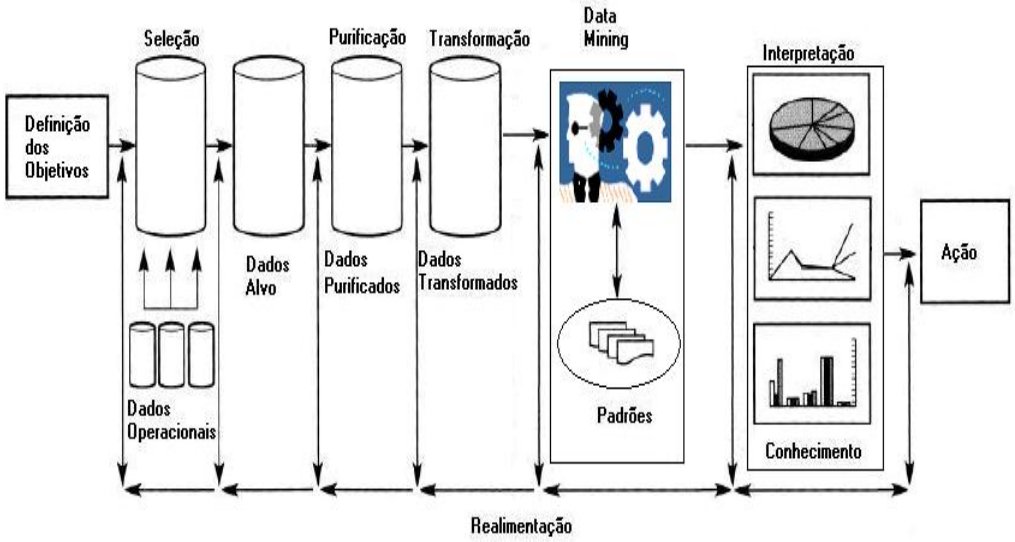


Figura 2.1: Etapas do DCBD adaptado de (FAYYAD, PIATETSKY-SHAPIRO & SMYTH, 1996).

2.1.2.1 - Definição dos objetivos

É nesta fase que as metas são traçadas, pois, para que um trabalho de descoberta de conhecimento tenha sucesso, é necessário estar claro o que se está buscando. Normalmente esta fase é feita com a ajuda de um especialista na área de aplicação.

2.1.2.2 - Seleção

Nesta fase seleciona-se um conjunto de dados ou focaliza-se um subconjunto de atributos ou de instâncias de dados, com objetivo de criar um conjunto de dados-alvo, no qual a descoberta será efetuada. Para realizar esta etapa, é necessário que se tenha uma compreensão do domínio e dos objetivos da tarefa, segundo AMARAL (2001, p. 15) e ROBIN e BEZERRA (2003).

2.1.2.3 - Purificação

Segundo HAN & KAMBER (2001, p. 109), nesta etapa é feita a limpeza dos dados, que envolve:

- o tratamento de campos de dados perdidos – que pode ser feito eliminando-se a tupla, ou usando, muitas vezes, médias dos valores presentes para preenchimento dos campos, dentre outros;
- redução ou eliminação de ruídos – que pode ser feito através de *binning* (substituir os valores ruidosos através de sorteio de valores pertencentes à vizinhança), de agrupamento ou inspeção humana com auxílio de ferramentas computacionais;
- correção de inconsistências nos dados – que trata da correção ou eliminação de dados inconsistentes. Um exemplo de dado inconsistente seria o atributo “cidade” ter os valores, “Florianópolis”, “Fpolis” ou “Floripa”.

2.1.2.4 - Transformação

De acordo com HAN & KAMBER (2001, p. 114), nesta etapa os dados são transformados de forma que se tornem apropriados à tarefa de mineração, à qual serão submetidas. Podendo envolver, dentre outros, a:

- agregação - muitas vezes não há necessidade de representar todas as faixas de valores de uma determinada variável. Pode-se reagrupá-las em faixas mais abrangentes, diminuindo assim o número de faixas de valores e a complexidade do problema;
- criação de atributos – em que atributos são criados e adicionados ao conjuntos de dados para auxiliar no processo de mineração;
- generalização dos dados – em que os valores iniciais (baixo nível) dos atributos são trocados por valores de alto nível no conceito hierárquico. Por exemplo, os valores do atributo “idade” podem ser substituídos por “jovem , adulto ou idoso”.

2.1.2.5 - Mineração de Dados

É nesta etapa que é feita a descoberta de conhecimento ou de padrões, propriamente dita. Neste momento, as técnicas são escolhidas de acordo com o tipo de problema a ser resolvido. Maiores detalhes sobre esta etapa serão apresentados na seção 2.2.

2.1.2.6 - Interpretação

Nesta etapa é feita a interpretação dos conhecimentos descobertos e o possível retorno aos passos anteriores. São removidos os padrões redundantes ou irrelevantes e traduzem-se os padrões úteis em termos compreensíveis aos usuários. Além disso, deve-se incorporar o conhecimento obtido, para melhorar o desempenho do sistema, adotando ações baseadas no conhecimento ou simplesmente documentando e relatando este conhecimento para grupos interessados.

2.2 – Mineração de Dados

Mineração de dados, de acordo com KLÖSGEN & ZYTKOW (2002, p. 2), é considerada a fase central do processo de DCBD. Esta fase é exclusivamente responsável pelo algoritmo minerador, ou seja, pelo algoritmo que diante da tarefa especificada busca extrair o conhecimento implícito e potencialmente útil dos dados.

2.2.1 –Fases da Mineração de dados

As fases da mineração de dados são cinco: escolha dos algoritmos a ser aplicada; descoberta de novas relações; análise humana das novas relações descobertas; uso racional das novas relações descobertas e avaliação dos resultados. A Figura 2.2 mostra essas fases, que são descritas a seguir.

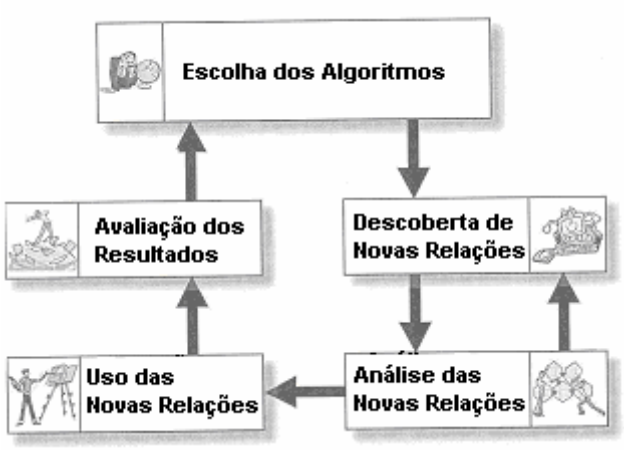


Figura 2.2: Fases do Mineração de dados , adaptado de CARVALHO (2001).

2.2.1.1 – Escolha dos Algoritmos

Nesta fase, os algoritmos são escolhidos de acordo com os objetivos definidos na fase inicial do DCBD, levando-se em consideração o tipo de dados que se tem para que a fase seguinte seja completada com êxito.

2.2.1.2 - Descoberta de Novas Relações

Nesta fase é que são descobertas novas relações que não são facilmente identificáveis, mas que podem ser visualizadas com a ajuda de algumas técnicas, por meio de uma análise sistemática e exaustiva sobre uma grande base de dados.

2.2.1.3 - Análise das Relações Descobertas

Nesta fase, as relações descobertas são analisadas por um especialista do domínio para verificar se tem algum valor informacional e se são coerentes. Deve-se também verificar se os objetivos foram atingidos totalmente, caso contrário deve-se voltar à fase anterior.

2.2.1.4 - Uso das relações descobertas

Nesta fase, as decisões são tomadas de forma a utilizar, da melhor maneira possível, as relações descobertas. A utilização dessas relações deve ser feita de forma racional para que se obtenha o melhor resultado possível.

2.2.1.5 - Avaliação dos Resultados

Esta é a fase final e é nela que se verifica se o problema foi resolvido ou se os objetivos foram alcançados. Por isso, ao começar um trabalho de DCBD, deve-se estar ciente de qual problema está se tentando resolver para que os resultados obtidos possam ser validados.

2.2.2 – Principais Técnicas da Mineração de dados

As principais técnicas de mineração de dados são: associação, classificação, agrupamento e análise de séries temporais, mostradas a seguir.

2.2.2.1 - Associação

Regras de associação são simples classes de sentenças que podem ser descobertas em grandes conjuntos de dados cujos valores são zeros e uns (zero para ausência de determinado acontecimento e um para presença). A sua utilidade reside na habilidade do algoritmo para encontrar todas as regras que satisfazem certas condições estabelecidas pelo usuário (KLÖSGEN & ZYTKOW, 2002, p. 344). Regras de associação são tratadas com mais detalhe na Seção 2.4.

2.2.2.2 - Classificação

De acordo com BERRY & LINOFF (1997, p. 52), classificação é uma tarefa que consiste em examinar características de um objeto e atribuí-lo a uma dentre várias classes pré-definidas. Segundo GROTH (2000, p. 22), classificação é o mapeamento de um conjunto de atributos para um conjunto de classes específicas.

Para BERRY & LINOFF (1997 p. 52), classificação é uma das técnicas mais utilizadas no processo de mineração, simplesmente porque é uma das tarefas cognitivas humanas mais realizadas no auxílio à compreensão do ambiente em que se vive. A mente humana naturalmente segmenta coisas em grupos distintos GROTH (2000, p. 22). O ser humano está sempre classificando as coisas ao seu redor: grupos de crianças, pessoas no trabalho, na escola, construções, por exemplo. GROTH (2000, cap. 6-7) mostra dois exemplos completos sobre classificação.

2.2.2.3 - Agrupamento

É um método que agrupa linhas de dados que compartilham tendências e padrões similares, ou seja, é o processo de dividir um conjunto de dados em grupos distintos (GROTH 2000, p. 247). De acordo com BARRY & LINOFF (1997, p. 55), é a tarefa de segmentar uma população heterogênea em um ou mais subgrupos homogêneos. E o que distingue agrupamento de classificação, é o fato de que no agrupamento não se tem grupos pré-definidos.

Cada grupo de objetos é formado de maneira que eles tenham alto grau de similaridade com outro objeto do mesmo grupo e alta dissimilaridade com objetos de outros grupos. Os grupos que são formados podem ser vistos como uma classe de objeto, da qual podem ser derivadas regras. Agrupamento também pode facilitar formação taxonômica, isto é, a organização de observações em uma hierarquia de classe (HAN & KAMBER, 2001, p.25).

2.2.2.4 - Análise de Séries Temporais (AST)

De acordo com BERK (1994, p. 321), AST tem a tarefa de analisar grandes conjuntos de dados de séries temporais para encontrar certas regularidades e características interessantes, incluindo busca por seqüência ou subseqüência similar e minerar padrões seqüenciais, periodicidades, tendências e desvios. Por exemplo, prever a quantidade de estoque para uma determinada época do ano para uma loja de departamento baseado em histórico do estoque, situação do negócio, desempenho dos concorrentes e mercado atual.

Análise de séries temporais é também uma busca por seqüência ou regras de seqüência que são, de acordo com NOTARI (2000), aquelas para as quais existe uma associação temporal nos fatos e, como nas regras de associação, existe um relacionamento de causa e efeito. A diferença é que nas regras de seqüência os itens que se relacionam estão em transações diferenciadas, ao contrário das regras de associação em quem os itens que se relacionam estão dentro da mesma transação.

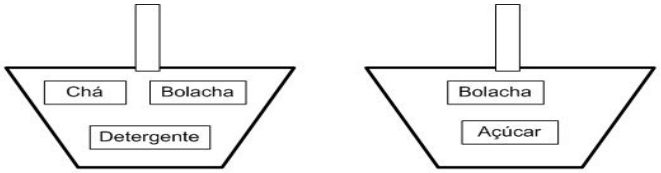
Regras de associação e *market basket analysis* são uns dos principais focos desta pesquisa. Dessa forma, torna-se necessária a apresentação mais detalhada sobre estes temas, o que será feita na próxima seção.

2.3 – Regras de Associação e Market Basket Analysis

Para GROTH (2000, p. 28), associação refere-se à informação comercial útil que pode ser extraída de associações agregadas entre os diferentes itens vendidos em catálogos ou em loja (física ou virtual). As entradas para a análise de associação são os dados transacionais dos pontos de vendas, e as saídas são informações e recomendações sobre associações entre produtos e comportamento de compra dos clientes.

De acordo com HAN & KAMBER (2001, p. 225), *Market Basket Analysis* (MBA), uma típica aplicação de regras de associação é o processo que analisa hábitos de compra de clientes para encontrar associações entre os diferentes itens que os clientes colocam em sua "cesta de compra". É uma técnica matemática, freqüentemente usada por profissionais de *marketing,* para revelar afinidades entre produtos individuais ou grupo de produtos. O nome *Market Basket Analysis* é uma analogia à idéia de que todos os clientes colocam suas compras em uma cesta.

O MBA é usado para determinar quais produtos são vendidos juntos, para o qual a entrada é normalmente uma lista de transações de vendas, e cada linha da tabela de dados representa uma venda ou um cliente, dependendo se o objetivo da análise é encontrar quais itens são vendidos juntos (ao mesmo tempo ou para o mesmo cliente). De acordo com GROTH (2000, p. 29), o MBA transforma dados transacionais em regras da forma "se cliente compra produto A, ele tende a comprar produto B, X% das vezes". Geralmente, os produtos são representados como atributos em uma base de dados. A Figura 2.3 mostra um exemplo de representação de produtos de duas cestas de mercado.



Cliente	Chá	Bolacha	Açúcar	Detergente
001	1	1	0	1
002	0	1	1	0

Figura 2.3: Representação de compra ou não compra de produtos em tabela de dados.

Como pode ser observado na Figura 2.3, cada item colocado na cesta representa um atributo ou coluna na tabela, e a existência do item em cada cesta é representada pelo valor lógico 1, e a não existência é representada pelo valor lógico 0. A cada produto diferente apresentado nas cestas, uma nova coluna na tabela é gerada, não importando quantos clientes compraram esse produto.

Segundo GROTH (2000, p. 29), MBA pode ser aplicado em vendas casadas, *layout* do mercado ou loja, projeto de catálogos de produtos, análise de perda de liderança, preço de produtos e promoções, dentre outros.

2.3.1 – O processo da MBA

O processo da MBA começa com a aquisição dos dados. A partir de uma base de dados de histórico de compra de clientes, como ilustrado na Figura 2.4, os dados são organizados em uma base de dados, como mostrado na Tabela 2.1.

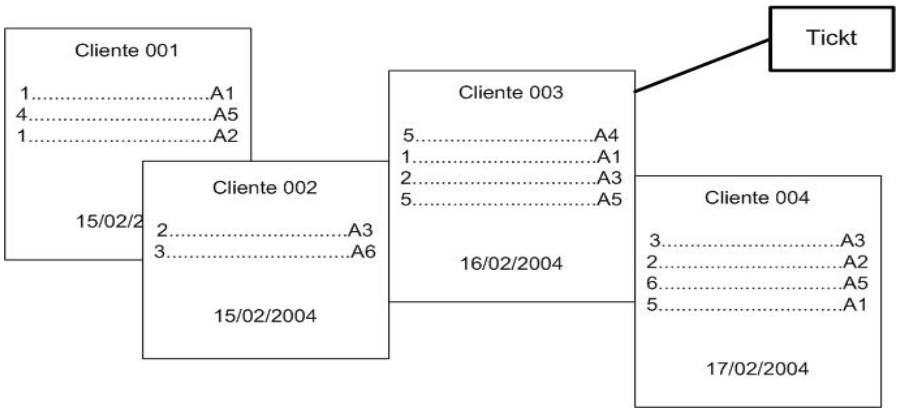


Figura 2.4: Base de dados de histórico de compra.

Tabela 2.1: Tabela de histórico de compra dos clientes, baseada em (AGRAWAL e SRIKANT, 1994)

Cliente	**Produto**				
001	A1	A2	A5		
002	A3	A5			
003	A1	A3	A4	A5	
004	A1	A2	A3	A5	
005	A1	A3	A4	A5	
006	A1	A1	A3	A4	A5
...	...	...	...	...	...

O processo de extração dos elementos da base de dados pode ser feito usando qualquer ferramenta de *query* (consulta). Na Tabela 2.1, os símbolos A1, A2,...,A5 representam os produtos comprados pelos clientes; como pode ser visto, essa forma de colocação dos dados, apesar de ser mais informativa do que estava na base de dados, ainda não é a ideal. Dessa forma, os dados ainda não estão prontos para o trabalho de MBA. Para que os dados estejam preparados para a tarefa de MBA, é necessário dispô-los da seguinte forma: os produtos se transformam em atributos e, quando está presente na cesta do cliente, a interseção do produto com o cliente recebe o valor lógico 1, caso contrário, recebe o valor lógico 0. Um exemplo é mostrado na Tabela 2.2, que também é conhecida como matriz de co-ocorrência.

Tabela 2.2: Matriz de co-ocorrência para a Tabela 2.1

Cliente	**A1**	**A2**	**A3**	**A4**	**A5**
001	1	1	0	0	1
002	0	0	1	0	1
003	1	0	1	1	1
004	1	1	1	0	1
005	1	0	1	1	1
006	1	1	1	1	1
...	...	...	...	...	...

Com os dados dispostos em uma matriz de co-ocorrência, começa-se o processo de extração de conhecimento da forma “se A então B”. Nesta fase calcula-se duas medidas: suporte e confiança que serão definidos a seguir.

2.3.1.1 - Suporte

É o percentual mínimo de transações na base de dados que contém os itens A e B (GROTH, 2000, p. 29), ou seja, o percentual dos casos em que a ocorrência de A “prevê” corretamente a ocorrência de B. Na prática, suporte é o percentual de ocorrência de A e B, simultaneamente, na matriz de co-ocorrência. O cálculo é simples, basta verificar a freqüência relativa de A e B, conforme (2.1). A Tabela 2.3 mostra o exemplo de um conjunto de dados para ilustrar o cálculo do suporte.

$$sup(A \rightarrow B) = prob(A\hat{}B) = \frac{n(A\hat{}B)}{N}, \qquad (2.1)$$

onde: $sup(A \rightarrow B)$ = suporte da regra “se A, então B”;

$prob(A\hat{}B)$ = probabilidade de ocorrência de A e B;

$n(A\hat{}B)$ = número de ocorrências simultâneas de A e B;

N = número de casos na base de dados.

Tabela 2.3: Dados para exemplificar o processo de MBA

Cliente	A1	A2	A3	A4	A5
001	1	1	0	1	0
002	1	1	1	0	0
003	0	0	1	0	1
004	1	0	1	0	1
005	0	1	1	1	1
006	0	1	0	1	1
007	0	1	0	1	1
008	1	1	0	0	0
009	0	0	1	0	1
010	1	0	1	1	0

Suponha-se que se queira calcular o suporte baseado na Tabela 2.3 para a regra “se o cliente compra A2 então compra A4”. O resultado seria 40% que significa que, se o cliente compra o produto A2, então ele compra o produto A4 em 40% dos casos. Poder-se-ia ter, por exemplo, mais de uma variável no antecedente do condicional, ou seja, “se o cliente comprar A1 e A3, então compra A5”. O resultado seria 10%, o que significa que em apenas 10% dos casos há ocorrências dos três itens simultaneamente.

2.3.1.2 - Confiança

Para (GROTH, 2000, p. 29), confiança é o percentual mínimo daquelas cestas que contém A e também contém B, ou seja, é o percentual dos cestos em que a co-ocorrência é observada. Em termos probabilísticos, confiança é a probabilidade de ocorrer B dado que A ocorreu, isto é, P(B|A) é a probabilidade condicional. O cálculo é feito da seguinte forma: basta dividir a freqüência relativa de A e B pela freqüência relativa de A, conforme (2.2).

$$conf(A \rightarrow B) = \frac{freq\_rel(A \wedge B)}{freq\_rel(A)}, \qquad (2.2)$$

onde: $conf(A \rightarrow B)$ = Confiança da regra “se A então B”,

$freq\_rel(A \wedge B)$ = Freqüência relativa de A e B,

$freq\_rel(A)$ = Freqüência relativa de A.

Suponha-se que se queira calcular a confiança baseada na Tabela 2.3 para a regra “se o cliente compra A2 então compra A4”. O resultado seria 66% que significa que esta regra aplica-se a 66% dos casos, ou seja, em 66% de todas as compras devem aparecer os produtos A2 e A4 juntos. Pode-se ter, por exemplo, mais de uma variável no antecedente do condicional, ou seja, “se o cliente comprar A1 e A3, então compra A5”. O resultado seria aproximadamente 33%, o que significa que se o cliente compra os produtos A1 e A3, ele compra o produto A5 em 33% dos casos.

Fazendo-se uma análise conjunta das duas medidas, para a regra “se o cliente comprar A1 e A3, então compra A5”, o conhecimento extraído da base do exemplo da

Tabela 2.3 seria: se o cliente comprar os produtos A1 e A3, então ele comprará o produto A5 em 33% dos casos e esta regra se aplicará a 10% dos casos.

GROTH (2000, p. 29-30) faz críticas a esses métodos de medida de associação, suporte e confiança, porque essas medidas não conseguem prever se a associação encontrada é casual ou não, ou se os produtos são concorrentes, isto é, em vez de se atraírem, se repelirem. Além disso, ele aponta problemas com a MBA, porque a maioria dos métodos não detecta se a associação entre os itens é casual ou não. Diz ainda que a medida de associação, confiança, é apenas uma probabilidade condicional de B visto que A ocorreu e, sozinho, não consegue distinguir entre uma associação casual ou uma associação útil. Com relação à medida de associação, suporte parece ser mais interessante, pois permite verificar transações pouco freqüentes.

Muitos trabalhos interessantes nessa área estão sendo desenvolvidos para que este tipo de problema seja resolvido. Pode-se encontrar na literatura algumas medidas para verificar se realmente há uma dependência entre os itens associados ou se a co-ocorrência dos itens não são casuais, como em (AUSLENDER, 2000), (BRIN et al., 1997), (BAKER et al., 2004) e (ARIA et al., 2002), entre outros. Essas medidas são apresentadas a seguir.

Para verificar se o item A é dependente de B e vice-versa, em uma associação do tipo “se A, então B”, calcula-se a co-ocorrência de A e B a priori e a posteriori. Se a diferença entre ambas for muito grande, pode-se considerar que A e B são dependentes e se ambas as probabilidades forem iguais ou aproximadamente iguais, A e B são independentes. Mas existe um problema em relação a essa medida de dependência. Como saber se a diferença é suficientemente grande? Para resolver tal problema, pode-se fazer o teste de significância estatística teste qui-quadrado, conforme (2.3). Para mais detalhe sobre o teste qui-quadrado, consulte BARBETTA (2004, p. 222-235).

$$\chi^2 = \frac{(freq\_relo(A \wedge B) - freq\_rele(A \wedge B))^2}{freq\_rRele(A \wedge B)}, \qquad (2.3)$$

onde: χ^2 = Medida de distância,

$freq\_rele(A \wedge B)$ = freqüência relativa esperada de A e B,

$freq\_relo(A \wedge B)$ = freqüência relativa obtida, de A e B.

Além desta, outras medidas podem ser encontradas, como em GROTH (2000, p. 87) que apresenta uma medida chamada impacto (*impact)* que é o quociente entre co-ocorrência obtida e a co-ocorrência esperada, conforme (2.4). Se o valor do impacto for próximo de 1, indica que os itens são independentes, caso contrário, são dependentes.

$$impact = \frac{freq\_rele(A \wedge B)}{freq\_relo(A \wedge B)}, \tag{2.4}$$

Outra medida muito comum, no mundo da MBA, é o *lift*, que GROTH (2000, p 87) apresenta conforme (2.5). O valor do *lift* é um valor que está entre –1 e 1. Caso este seja igual a 0, A e B são independentes; caso seja negativo, A e B se repelem; caso seja positivo, A e B se atraem.

$$lift(A \rightarrow B) = \frac{freq\_relo(A \wedge B) - freq\_rele(A \wedge B)}{freq\_rel(A)}, \tag{2.5}$$

onde: $lift(A \rightarrow B)$ = Medida de atração entre os itens A e B,

$freq\_rele(A \wedge B)$ = Freqüência Relativa esperada de A e B,

$freq\_relo(A \wedge B)$ = Freqüência Relativa obtida de A e B,

f*req_rel(A)* = Freqüência Relativa de A.

BERRY& LINOFF (1997, p.107-109) apresentam outra medida, “informatividade” da associação “se A, então B”, cujo nome é também *lift* e é calculada conforme (2.6).

$$lift(A \rightarrow B) = \frac{conf(A \rightarrow B)}{freq\_rel(B)} = \frac{\frac{freq\_rel(A\hat{}B)}{freq\_rel(A)}}{freq\_rel(B)} = \frac{freq\_rel(A\hat{}B)}{freq\_rel(A) * freq\_rel(B)}, \tag{2.6}$$

onde: $lift(A \rightarrow B)$ = medida de atração/repulsão da associação se A, então B,

$conf(A \rightarrow B)$ = Confiança da associação A, então B,

$freq\_rel(B)$ = Freqüência relativa de B.

A faixa de valores das equações (2.5) e (2.6) é diferente apesar de ambas as equações terem o mesmo propósito: verificar se A e B são dependentes ou independentes, ou se repelem, ou se atraem. A faixa de valores de (2.5) varia de –1 a 1, onde um valor no intervalo [-1, 0) indica repulsão; um valor igual a 0 indica independência e um valor no intervalo (0, 1] indica atração. Já (2.6) só possui o limite inferior definido igual a 0, pois o limite superior depende do valor do denominador de (2.6). Por exemplo, se o valor do denominador for igual a 0,1, o limite superior será igual a 10. A interpretação se dá da seguinte forma: um valor menor que 1 indica associação negativa, ou seja, repulsão; um valor igual a 1 indica independência, e um valor maior do que 1 indica associação positiva, ou seja, atração.

Outras medidas de associação podem ser encontradas, que são: *coverage* e *leverage* que são apresentados, respectivamente, pelas equações (2.7) e (2.8). *Coverage* indica a proporção de exemplos no conjunto de dados que é coberto pelo antecedente da regra, e *leverage* é a medida de importância da associação que é refletida pela cobertura e confiança.

$$cov(A \rightarrow B) = freq\_relo(A), \tag{2.7}$$

onde: $cov(A \rightarrow B)$ = *Coverage* para a regra se A, então B,

$freq\_rel(A)$ = Freqüência relativa de A.

$$lev(A \rightarrow B) = sup(A \rightarrow B) - freq\_rel(A) * freq\_rel(B) = \frac{n(A \wedge B)}{N} - freq\_rel(A) * freq\_rel(B), \tag{2.8}$$

onde: $lev(A \rightarrow B)$ = *leverage* para a regra se A, então B,

$sup(A \rightarrow B)$ = suporte para a regra se A, então B, calculado conforme (2.1).

Deste estudo elaborou-se um resumo das medidas de associação apresentando-se a Tabela 2.4.

Tabela 2.4: Resumo das medidas de associação usadas em MBA

Medida	Equação	Descrição
Suporte	$sup(A \rightarrow B) = prob(A\hat{}B) = \frac{n(A\hat{}B)}{N}$	É o percentual mínimo de transações na base de dados que contém os itens A e B.
Confiança	$conf(A \rightarrow B) = \frac{freq\_rel(A \wedge B)}{freq\_rel(A)}$	É o percentual mínimo daquelas cestas que contém A e também contém B.
Qui-Quadrado	$\chi^2 = \frac{(freq\_relo(A \wedge B) - freq\_rele(A \wedge B))^2}{freq\_rRele(A \wedge B)}$	Medida de distância entre a freqüência esperada e obtida, de A e B.
Impact	$impact = \frac{freq\_rele(A \wedge B)}{freq\_relo(A \wedge B)}$	Se o valor do impacto for próximo de 1, indica que os itens são independentes, caso contrário, são dependentes.
Lift (Groth, 2000)	$lift(A \rightarrow B) = \frac{freq\_relo(A \wedge B) - freq\_rele(A \wedge B)}{freq\_rel(A)}$	Verifica se A e B são dependentes ou independentes e se repelem ou se atraem.
Lift BERRY& LINOFF (1997, p.107-109)	$lift(A \rightarrow B) = \frac{conf(A \rightarrow B)}{freq\_rel(B)}$	Mede a "informatividade" da associação, ou seja, indica quão freqüente é B em relação a A.
Coverage	$cov(A \rightarrow B) = freq\_relo(A)$	Indica a proporção de exemplos no conjunto de dados que é coberto pelo antecedente da regra.
Leverage	$lev(A \rightarrow B) = sup(A \rightarrow B) - freq\_rel(A) * freq\_rel(B)$	É a medida de importância da associação que é refletida pela cobertura e confiança.

Como pôde ser visto nesta seção, existem várias medidas para quantificar a associação entre A e B, conforme resumo apresentado na Tabela 2.4. Porém, todas elas exigem uma interpretação que não pode ser feita por usuários não especialistas. Exige-se que o usuário conheça no mínimo um pouco de estatística, para que possa tirar proveito do conhecimento extraído da base de dados. O ideal seria uma medida que pudesse dizer ao usuário o que esse conhecimento significa. Para tanto, foi necessário desenvolver um método que trabalha com imprecisão devido ao fato de que a comunicação com o usuário exige o entendimento do raciocínio humano, que é impreciso. Como a proposta desta pesquisa foi verificar a adequação da lógica difusa, que é baseada na teoria dos conjuntos difusos, foi feito um estudo sobre estes dois tópicos que é apresentado na próxima seção.

2.4 – Lógica Difusa

Nesta seção são apresentados conceitos sobre lógica difusa. Porém, inicialmente, é feita uma breve introdução à teoria da lógica clássica.

De acordo com o “postulado do meio excluído” da lógica clássica, toda proposição só admite o valor verdadeiro ou falso, não existindo um termo intermediário. Segundo BARRETO (2000, p. 42), os profissionais da computação encontram nesse postulado correspondência natural com os dois estados dos circuitos empregados nos sistemas computacionais. Porém, as proposições no mundo real admitem valores diferentes desses dois estados, e a limitação da lógica clássica dificulta expressar, com precisão satisfatória, as idéias cujas origens são as sensações e percepções humanas.

Existem lógicas que admitem valor intermediário, são as lógicas multi-valoradas. As lógicas multi-valoradas mais conhecidas são a de Kleene (tri-valorada) que considera um terceiro valor como ignorância, ou seja, um valor que pode assumir tanto verdadeiro como falso; e a de Lukasievsky que, de acordo com BARRETO (2000, p. 42), considera a possibilidade de tantos valores intermediários quantos forem necessários para melhor captar os níveis de verdade possíveis, sendo um passo para a lógica difusa, que é definida a seguir.

Lógica difusa trabalha com algoritmos usados para emular pensamentos e decisões humanas em máquinas. É a lógica que serve de base para os modos de raciocínio que são aproximados, ao invés de exatos (TANSCHEIT, 2003, p. 1).

De acordo com BARRETO (2000, p. 42), assim como a lógica de primeira ordem tem sua correspondente na teoria dos conjuntos clássicos, a lógica difusa tem sua correspondente na teoria dos conjuntos difusos. Para melhor compreensão destes faz-se a seguir uma apresentação dos conjuntos clássicos.

2.4.1 Conjuntos Clássicos

Dado um conjunto clássico *A*, em um universo de discurso U, pode-se definir *A* listando todos os seus elementos ou através de uma propriedade que os identifica (MENDEL, 1995, p. 348). Pode-se definir uma função para indicar se um determinado elemento pertence ou não a um conjunto, sendo atribuído o valor zero para não pertinência $x \notin A$ e o valor um para pertinência $x \in A$. Essa função é chamada por MENDEL (1995, p. 348) de função zero-um ou função característica.

$$\mu_A(x) : U \rightarrow \{0,1\} \tag{2.9}$$

onde,

$$\mu_A(x) = \begin{cases} 1, & se \quad e \quad somente \quad se \quad x \in A \\ 0, & se \quad e \quad somente \quad se \quad x \notin A \end{cases} \tag{2.10}$$

De acordo com SCREMIN (2003, p. 46), a teoria dos conjuntos clássicos admite apenas resultados binários 0 ou 1 quanto à pertinência de um elemento a um conjunto não permite que esse elemento pertença parcialmente a outro conjunto. Para mais detalhes sobre lógica clássica, consulte NOLT (1991), DEVLIN (1991), GABBAY (1994) e NISSANKE (1999).

Em meados dos anos 60, Lofti Zadeh (ZADEH, 1965) desenvolveu a teoria de conjuntos difusos que permite trabalhar de acordo com o raciocínio humano, que é intrinsecamente impreciso e vago. Essa teoria diz que um conjunto não apresenta necessariamente limites bem definidos, podendo um elemento pertencer parcialmente a um conjunto ou pertencer a dois conjuntos ao mesmo tempo.

2.4.2 Conjuntos Difusos

Nesta seção, são apresentados os conceitos essenciais sobre conjuntos difusos para o desenvolvimento desta pesquisa. Todos os conceitos e definições são baseados em KANDEL (1986), KLIR (1995), KOSKO (1997), ROSS (1995) e YAGER (1987).

2.4.2.1 Conceitos e Definições

Definição de conjuntos difusos

Seja X um conjunto de pontos com um elemento genérico denotado por x, assim $x \in X$. Um conjunto difuso $A \subset X$ é caracterizado por uma função característica $\mu_A(x)$ que associa cada elemento de A a um número real em um intervalo [0,1], onde $\mu_A(x)$ representa o grau de pertinência de x ao conjunto A.

A função $\mu_A(x)$ é referida como função de pertinência ou função de associação e é representada matematicamente conforme (2.11).

$$\mu_A(x) : X \rightarrow [0,\ 1]\ , \tag{2.11}$$

onde $\mu_A(x)$ representa o grau de pertinência de x ao conjunto A.

A representação algébrica mais usada de um conjunto difuso é apresentada pelas equações (2.12) e (2.13), que representam, respectivamente, os conjuntos difusos, discreto e contínuo.

$$A = \left\{ \frac{\mu_A(x_1)}{x_1} + \frac{\mu_A(x_2)}{x_2} + ... + \frac{\mu_A(x_n)}{x_n} \right\} = \sum_{i=1}^{n} \frac{\mu_A(x_i)}{x_i} \tag{2.12}$$

$$A = \int_x \frac{\mu_A(x)}{x} dx\,, \tag{2.13}$$

onde o sinal (+) representa a operação união e não uma soma aritmética.

Suporte

O suporte de A é o conjunto de pontos em X tal que $\mu_A(x) > 0$, que pode ser escrito conforme (2.14).

$$Sup(A) = \{x \in X, \mu_A(x) > 0\} \tag{2.14}$$

***Alfa-Cut* ou Corte Alfa** ($\alpha - cut$)

Dado um conjunto difuso A definido em X e algum número $\alpha \in [0,1]$, o corte α ${}^{\alpha}A$, e o corte α forte (que é uma variante do corte α) ${}^{\alpha+}A$ são os conjuntos clássicos definidos pelas equações (2.15) e (2.16), respectivamente.

$$ {}^{\alpha}A = \{x \mid A(x) \geq \alpha\} \tag{2.15} $$

e

$$ {}^{\alpha+}A = \{x \mid A(x) > \alpha\}, \tag{2.16} $$

onde o $\alpha\, cut$ ou corte α forte de um conjunto difuso A é o conjunto clássico ${}^{\alpha}A$ ou o conjunto clássico ${}^{\alpha+}A$ que contém todos os elementos do conjunto universo X cujos graus de pertinência em A são, respectivamente, maiores ou iguais, ou somente maior do que um valor específico α .

Principais Operações com Conjuntos Difusos

As operações básicas da teoria dos conjuntos difusos são: complemento, interseção (t-norma) e união (t-conorma) e são baseadas no conceito de pertinência ou não de um elemento a um conjunto difuso. Há diferentes propostas para encontrar suas funções de pertinência.

A função de pertinência do complemento padrão de um conjunto difuso A, com função de pertinência μ_A, é definida como o complemento da correspondente função de pertinência, também chamado de critério da negação; ela é representada matematicamente através de (2.17).

$$ \mu_{\bar{A}} = 1 - \mu_A \tag{2.17} $$

A função de pertinência da interseção padrão de dois conjuntos difusos A e B, com funções de pertinência μ_A e μ_B, é definida como o mínimo das duas funções de

pertinência individuais, também chamada de critério dos mínimos; ela é representada matematicamente conforme (2.18).

$$\mu_S[A \cap B(x)] = min[\mu_A(x),\mu_B(x)] \tag{2.18}$$

Além da t-norma, interseção padrão, apresentada em (2.18), existem outras t normas, das quais podem-se citar diferença limitada, produto algébrico e interseção drástica que são representadas matematicamente conforme (2.19), (2.20) e (2.21), respectivamente.

$$\mu_b[A \cap B(x)] = Max[0, \mu_A(x) + \mu_B(x) - 1] \tag{2.19}$$

$$\mu_p[A \cap B(x)] = Max[\mu_A(x).\mu_B(x)] \tag{2.20}$$

$$\mu_d[A \cap B(x)] = \begin{cases} \mu_A(x), & se \quad \mu_B(x) = 1 \\ \mu_B(x), & se \quad \mu_A(x) = 1 \\ 0, & caso \quad contrário \end{cases} \tag{2.21}$$

A função de pertinência da união padrão de dois conjuntos difusos *A* e *B*, com funções de pertinência μ_A e μ_B, é definida como o máximo das duas funções de pertinência individuais, também chamada de critério dos máximos ou t-conorma; ela é representada matematicamente conforme (2.22).

$$\mu_{A \cup B}(x) = Max(\mu_A(x), \mu_B(x)) \tag{2.22}$$

Assim como a t-norma, existem outras t-conormas além da apresentada conforme (2.22) das quais podem-se citar soma limitada, soma-produto e união drástica que são representadas matematicamente conforme (2.23), (2.24) e (2.25), respectivamente.

$$\mu_b[A \cup B(x)] = Min[1, \mu_A(x) + \mu_B(x) + \mu_A(x).\mu_B(x)] \tag{2.23}$$

$$\mu_p[A \cup B(x)] = Min[\mu_A(x) + \mu_B(x) - \mu_A(x).\mu_B(x)] \tag{2.24}$$

$$\mu_d[A \cup B(x)] = \begin{cases} \mu_A(x), & se \quad \mu_B(x) = 0 \\ \mu_B(x), & se \quad \mu_A(x) = 0 \\ 1, & caso \quad contrário \end{cases} \tag{2.25}$$

2.4.2.2 Funções de pertinência

Nesta seção são apresentadas as principais funções de pertinência e suas representações gráficas. As funções de pertinência podem ser lineares ou não-lineares. As lineares apresentadas aqui são a triangular, a trapezoidal a gama a L, e as não lineares são Gaussiana, Z e sigmoidal.

Função triangular

Seja $a,b,c \in \Re$ (conjunto dos números reais), define-se a função triangular conforme (2.26). E sua representação gráfica é ilustrada pela Figura 2.5.

$$\mu(x) = \begin{cases} \dfrac{x-a}{b-a}, & se \quad a \le x \le b \\ \dfrac{c-x}{c-b}, & se \quad b \le x \le c \\ 0, & caso \quad contrário \end{cases} \tag{2.26}$$

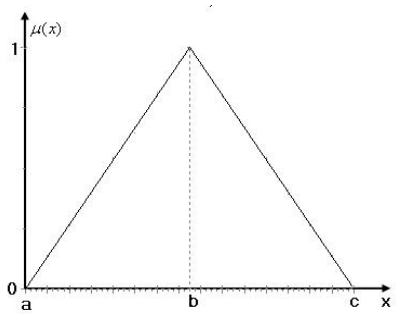


Figura 2.5: Função de pertinência de formato triangular.

Função trapezoidal

Seja $a,b,c,d \in \Re$, define-se a função trapezoidal conforme (2.27). Sua representação gráfica é ilustrada pela Figura 2.6.

$$\mu(x) = \begin{cases} 0, & se \quad x \le a \\ \dfrac{x-a}{b-a}, & se \quad a < x \le b \\ 1, & se \quad b < x \le c \\ \dfrac{d-x}{d-c}, & se \quad c < x \le d \\ 0, & se \quad x > d \end{cases} \tag{2.27}$$

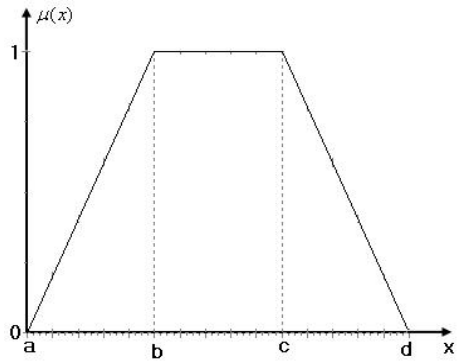


Figura 2.6: Função de pertinência de formato trapezoidal.

Função gama

Seja $a,b,c \in \Re$, define-se a função gama conforme (2.28). Sua representação gráfica é ilustrada pela Figura 2.7.

$$\mu(x) = \begin{cases} 0, & se \quad x \le a \\ \dfrac{x-a}{b-a} & se \quad a < x \le b \\ 1, & se \quad b < x \le c \end{cases} \tag{2.28}$$

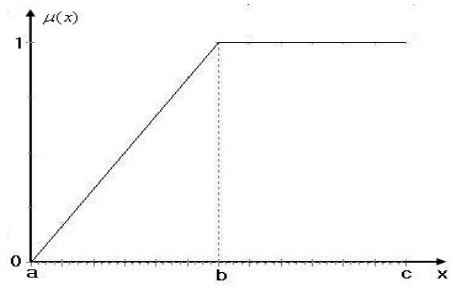


Figura 2.7: Função de pertinência do tipo gama.

Função L

Seja $a,b,c \in \Re$, define-se a função L conforme (2.29). Sua representação gráfica é ilustrada pela Figura 2.8.

$$\mu(x) = \begin{cases} 1, & se \quad a \le x < b \\ \dfrac{c-x}{c-b} & se \quad b < x \le c \\ 0, & se \quad x > c \end{cases} \tag{2.29}$$

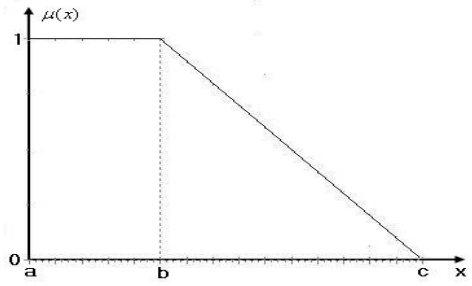


Figura 2.8: Função de pertinência do tipo L.

Função Gaussiana

A função Gaussiana, também chamada de π, é definida por (2.30), onde α é o valor de x para qual a função $\mu(x)$ tem ponto de máximo, e β representa a largura do intervalo, onde $\mu(x) = 0$. Sua representação gráfica é ilustrada pela Figura 2.9.

$$\mu(x) = \begin{cases} 0, & se \quad x \le \alpha - \beta \, e \, x \ge \alpha + \beta \\ \frac{2}{\beta^2}(x - \alpha + \beta)^2, & se \quad \alpha - \beta \le x \le \alpha - \frac{\beta}{2} \\ 1 - \frac{2}{\beta^2}(x - \alpha)^2, & se \quad \alpha - \frac{\beta}{2} \le x \alpha + \frac{\beta}{2} \\ \frac{2}{\beta^2}(x - \alpha - \beta)^2, & se \quad \alpha + \frac{\beta}{2} \le x \le \alpha + \beta \end{cases} \tag{2.30}$$

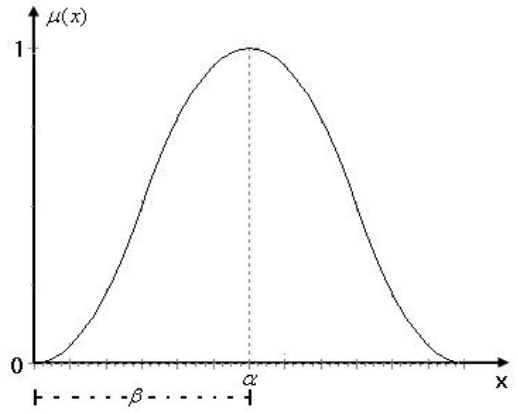


Figura 2.9: Função de pertinência de formato π.

Função Z

A função Z é definida por (2.31), onde α é o valor de *x* para qual $\mu(x)$ tem um ponto de máximo, e β representa a largura do intervalo. Sua representação gráfica é ilustrada pela Figura 2.10.

$$\mu(x) = \begin{cases} 1, & se \quad x \le \alpha \\ \frac{2}{\beta^2}(x - \alpha - \beta)^2, & se \quad \alpha - \frac{\beta}{2} \le x \le \alpha + \frac{\beta}{2} \\ 1 - \frac{2}{\beta^2}(x - \alpha)^2, & se \quad \alpha + \frac{\beta}{2} \le x \le \alpha + \beta \\ 0, & se \quad x \ge \alpha + \beta \end{cases} \tag{2.31}$$

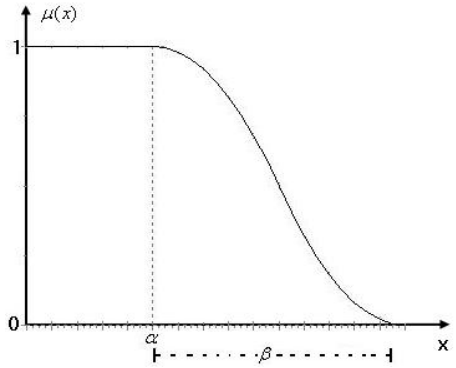


Figura 2.10: Função de pertinência de formato Z.

Função Sigmoidal

A função sigmoidal é definida por (2.32), onde α é a inclinação no ponto de transição, e β define o ponto de transição. Sua representação gráfica é ilustrada pela Figura 2.11.

$$\mu(x) = \begin{cases} 0, & se \quad x \le \alpha - \beta \\ \frac{2}{\beta^2}(x - \alpha + \beta)^2, & se \quad \alpha - \beta \le x \le \alpha - \frac{\beta}{2} \\ 1 - \frac{2}{\beta^2}(x - \alpha)^2, & se \quad \alpha - \frac{\beta}{2} \le x\alpha \\ 1, & se \quad x \ge \alpha \end{cases} \tag{2.32}$$

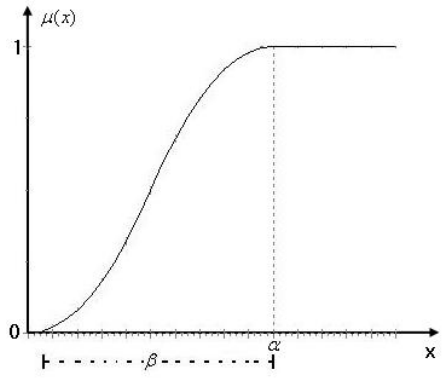


Figura 2.11: Função de pertinência sigmoidal.

Além dessas outras funções de pertinência podem ser encontradas na literatura, como é o caso da função sino (Figura B.8), função gama e função L, dentre outras. Às vezes, a mesma função pode aparecer com nomes diferentes em diferentes textos, como é o caso da função *sigmoidal* que pode ser encontrada com o nome de função *S.* Existem combinações de funções, como é o caso da função *sino-sigmóide*.

2.4.2.3 Modelos de Regras Difusas

As regras difusas formam a parte fundamental da estrutura de conhecimento em um sistema difuso de inferência. Os formatos de regras difusas podem ser classificados em quatro grupos principais. Os três primeiros, *Mamdani* e *Larsen*, *Takagi-Sugeno* e *Tsukamoto* correspondem ao modelo de inferência difuso. A diferença básica entre esses três primeiros modelos recai no tipo de conseqüente e no procedimento de desfuzificação. A seguir são descritos os quatro formatos.

Modelo *Mamdani* e *Larsen*

No modelo *Mamdani* e *Larsen,* as regras são do tipo: *Rj: se* x_1 *é* $A_{1,j}$ *e...e* x_n *é* $A_{n,j}$ então y_j é C_j. Sendo que no modelo de *Mamdani* a saída é obtida através do seguinte processo: calcula-se, inicialmente, as t-normas usando interseção padrão, conforme (2.18); em seguida, calcula-se o mínimo entre as t-normas para cada regra disparada; finalmente, a saída é o máximo entre esses mínimos. Esse processo, ilustrado pela Figura 2.12, é conhecido como Max-Min. Já no modelo de *Larsen*, em vez de interseção padrão, calcula-se as t-normas usando produto algébrico, conforme (2.20); em seguida, obtém-se o produto das t-normas para cada regra disparada e, para se obter a saída final, procede-se da mesma forma do que no modelo de *Mamdani*. Este processo, ilustrado pela Figura 2.13, é conhecido como Max-Prod.

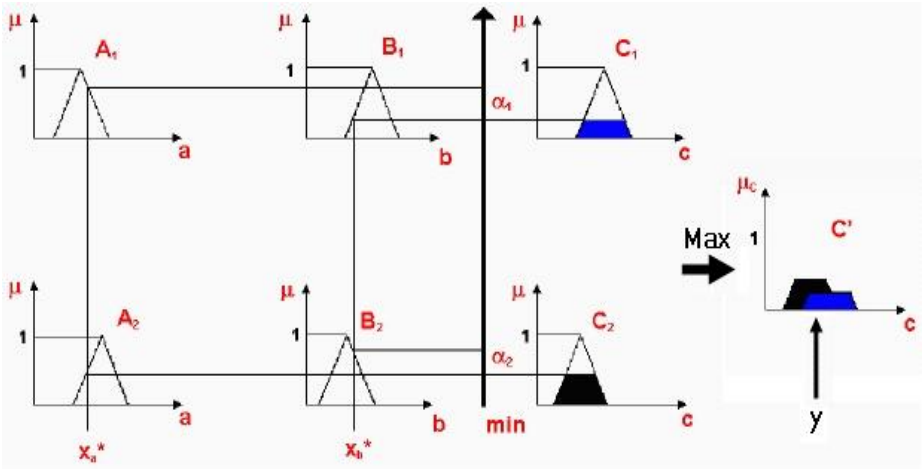


Figura 2.12: Modelo de *Mamdani* com composição Max-Min adaptado de (SANDRI & CORREA, 1999, p. c80).

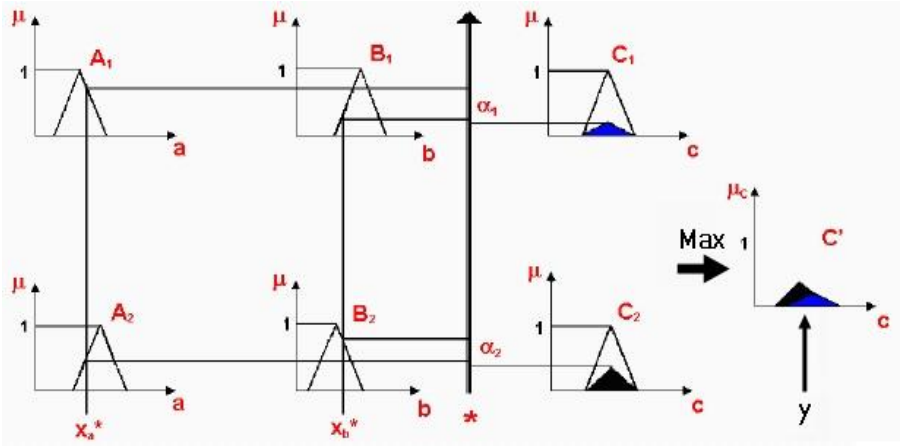


Figura 2.13: Modelo de *Larsen* com composição Max-Prod adaptado de (SANDRI & CORREA, 1999, p. c80).

Modelo de *Takagi-Sugeno*

No modelo *Takagi-Sugeno,* as regras são do tipo: *Rj: se* x_1 *é* $A_{1,j}$ *e...e* x_n *é* $A_{n,j}$ *então* $y_j = f_j(x_1,...,x_m)$. Nesse caso, a saída de cada regra é uma função das variáveis de entrada. Geralmente, a função que mapeia a entrada e saída para cada regra é uma combinação linear das entradas, isto é, $y = d_0 + d_{1,j}x_1 + ... + d_{m,j}x_m$. No caso em que $d_1 = ... = d_m = 0$, tem-se y = d_0 (*fuzzy singleton*). A saída do sistema é obtida pela média ponderada, procedimento de desfuzificação, das saídas de cada regra usando-se o grau de disparo dessas regras como pesos de ponderação. A Figura 2.14 ilustra esse processo.

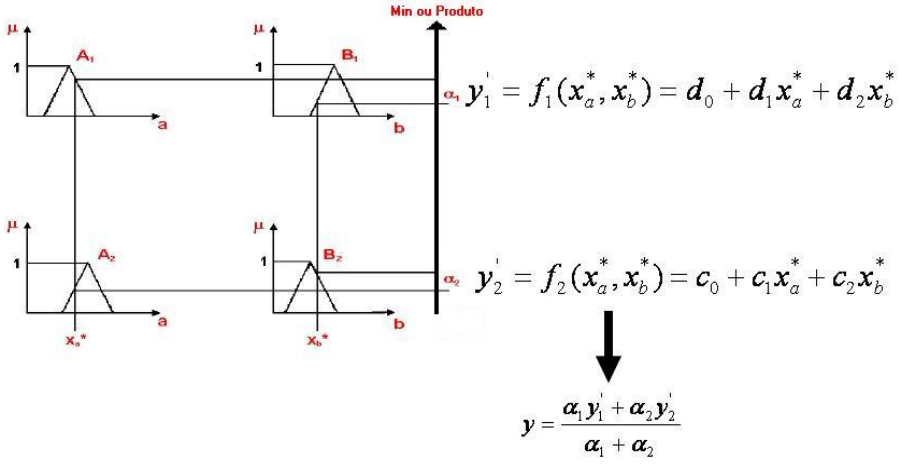


Figura 2.14: Modelo de Takagi-Sugeno adaptado de (SANDRI & CORREA, 1999, p. c81).

Modelo de Tsukamoto

No modelo *Tsukamoto,* as regras são do tipo: *Rj: se* x_1 *é* $A_{1,j}$ *e...e* x_n é $\mathbf{A_{n,j}}$ então y_j é C_j. Nesse sistema difuso de inferência o conseqüente de cada regra é representado por um conjunto difuso com uma função de pertinência monotônica, conforme ilustrado pela Figura 2.15.

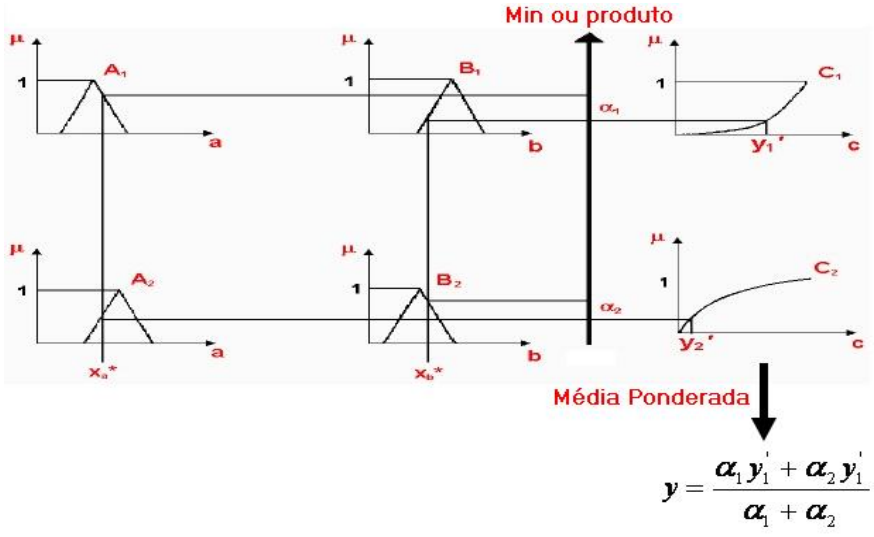


Figura 2.15: Modelo de *Tsukamoto* adaptado de (SANDRI & CORREA, 1999, p. c80).

Modelo para classificação

Neste modelo, as regras são do tipo: *Se x é A e y é B então padrão (x,y) pertence a classe i*. Segundo SOUZA (1999, p. 14), este modelo de regras difuso foi acrescentado aos modelos clássicos anteriores pelo fato dos demais não serem adequados aos sistemas de inferência difusa desenvolvidos para tarefas de classificação. A Figura 2.16 ilustra um exemplo de um sistema difuso para classificação com duas entradas e três classes de saída. Nesse modelo, as saídas são calculadas diretamente pelas operações de t-conorma aplicadas sobre o grau de disparo das regras (t-normas). Nesse caso não há procedimento de desfuzificação.

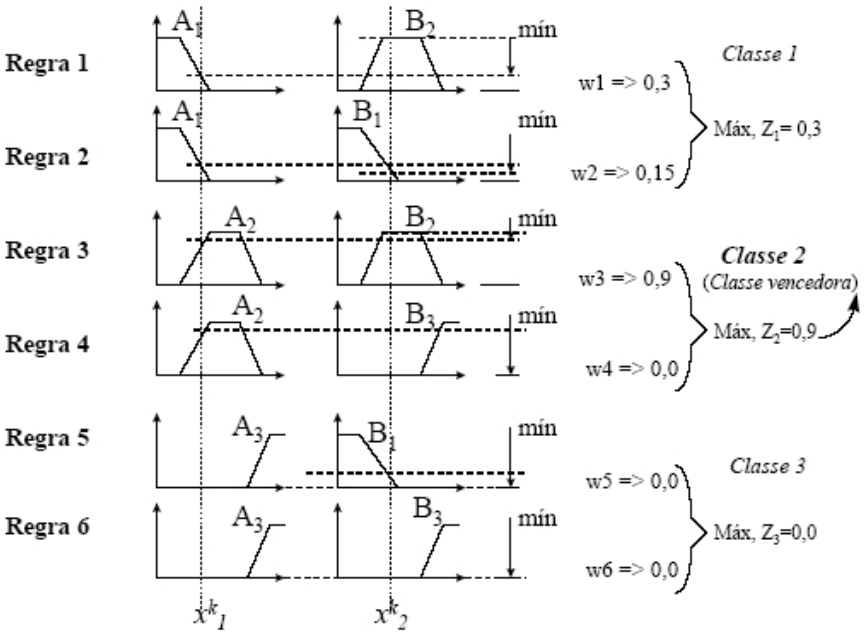


Figura 2.16: Modelo *fuzzy* de classificação com duas entradas e três classes de saída (SOUZA, 1999).

As variáveis lingüísticas são partes integrantes de um sistema difuso, pois elas são responsáveis por representar termos lingüísticos. Para melhor entendimento do seu significado, a próxima seção apresenta a definição formal e um exemplo.

2.4.2.4 Variáveis Lingüísticas

Uma variável lingüística difusa é uma variável cujo valor é expresso qualitativamente por um termo lingüístico e quantitativamente por uma função de pertinência. Uma variável lingüística é caracterizada pela t-upla *{n,T,X,m(t)},* onde *n* é o nome da variável, como, por exemplo, temperatura, pressão, febre, sabor e dor, *T* é o conjunto de termos lingüísticos de *n*, como, por exemplo, elevada, baixa, extrema, suave e intensa, X é o domínio de valores de *n* sobre o qual o significado do termo lingüístico é determinado e *m(t)* é uma função semântica que assinala para cada termo lingüístico $t \in T$ o seu significado, que é um conjunto difuso em *X*, ou seja, $m : T \rightarrow X$, onde *X* é o espaço dos conjuntos difusos.

A Figura 2.17 mostra um exemplo de variável lingüística. O nome *(n)* da variável é FRVG (Freqüência Relativa de Venda de Geladeira). Os termos lingüísticos $t \in T$ que atribuem um significado semi-quantitativo a FRVG são baixa, moderada e alta. O domínio *X* da variável é o intervalo [0,1]. Cada termo lingüístico tem a ele associado um conjunto difuso *m(t)* que o caracteriza.

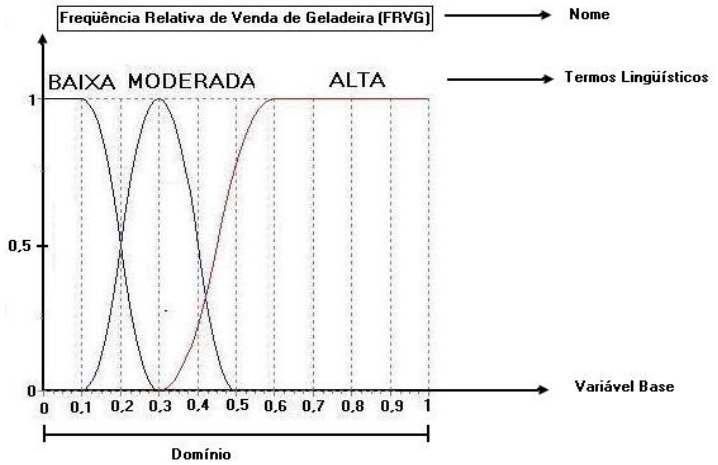


Figura 2.17: Exemplo de variáveis lingüísticas.

Os conceitos apresentados nesta seção têm como objetivo dar uma idéia geral sobre a teoria da lógica difusa, enfocando apenas os tópicos que dão suporte a esta pesquisa.

Neste capítulo foram apresentados conceitos de Descoberta de Conhecimento em Base de Dados (DCBD) dando destaque a uma de suas etapas, a mineração de dados. A mineração de dados tem um grande papel no processo de DCBD, porque é nesta etapa que são escolhidas e aplicadas as técnicas de mineração de dados para a descoberta do conhecimento. Uma das técnicas é a descoberta de regras de associação que é bastante usada no processo de *Market Basket Analysis* (MBA), que também foi discutido. O capítulo se fecha com uma revisão da teoria sobre conjunto difuso e também sobre lógica difusa, assuntos que serviram para completar o embasamento teórico que deu suporte a esta pesquisa, visto que se trata de uma investigação do uso de conjunto difuso para tratamento de imprecisão contida na matriz de co-ocorrência utilizada no processo de MBA.

Capítulo 3 – Método Difuso para Cálculo de Atração e Repulsão (MDCAR)

Devido ao fato de as medidas utilizadas no cálculo de atração e repulsão entre itens não tratarem a imprecisão contida na matriz de co-ocorrência, procurou-se desenvolver um método baseado na teoria de conjuntos difusos e na lógica difusa para resolver este problema. Para isso, procurou-se mapear as medidas de atração/repulsão mais utilizadas que são *lift* de Berry-Linoff (BERRY e LINOFF, 1997) e *lifit* de Groth (GROTH, 2000), ou seja, usar essas duas medidas como entrada. Mas não foi possível usa-las diretamente visto que a de Berry-Linoff não tem um limite superior definido como descrito na Seção 2.3.1. Por isso foi necessário fazer algumas adaptações nestas medidas, o que levou às seguintes entradas FRA (Freqüência Relativa do antecedente da regra “se A então B”), FREAB (Freqüência Relativa Esperada do antecedente e conseqüente da regra “se A então B”) e FROAB (Freqüência Relativa Obtida do antecedente e conseqüente da regra “se A então B”). Essas entradas são valores numéricos percentuais que são mapeadas para uma saída qualitativa conforme ilustrado pela Figura 3.1 e descritas na Seção 3.1.1. A próxima seção apresenta uma descrição detalhada do método MDCAR.

3.1 – Descrição do método

A Figura 3.1 mostra o esquema geral da proposta do método difuso para cálculo de atração e repulsão, o MDCAR, que começa com a fuzificação das entradas, terminando com o fornecimento de uma saída qualitativa do MDCAR. Cada etapa do processo é descrita a seguir.

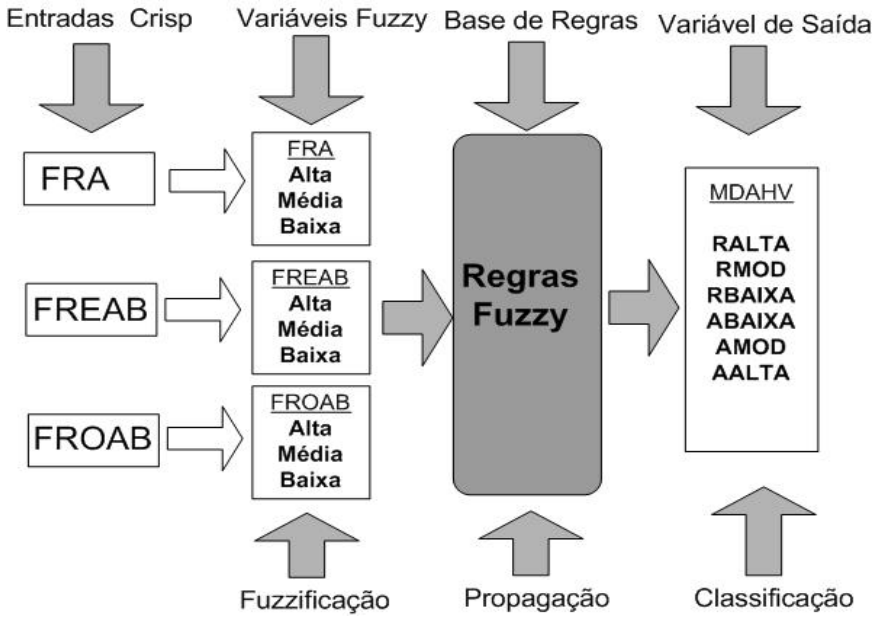


Figura 3.1: Esquema do MDCAR.

3.1.1 - Entradas Numéricas Percentual

As entradas numéricas, cujos valores variam entre 0 e 1, são as mesmas usadas para calcular o *lift* de Groth (GROTH, 2000, p. 87) (2.5) que é a medida usada como referência para desenvolvimento do método proposto, que são:

- FRA - Freqüência Relativa de A, obtida por (3.1), onde A é o antecedente da regra do tipo "Se A, então B", que pode ser, por exemplo, "Se o cliente compra o item P_x, ou os itens P_1, P_2, ..., P_n, então compra também o item Q_y, ou produtos Q_1, Q_2, ..., Q_n;

- FREAB – Freqüência Relativa Esperada de A e B, onde B é o conseqüente da regra descrita anteriormente, obtida por (3.2);

- FROAB – Freqüência Relativa Obtida de A e B, que é calculada por (3.4).

$$FRA = freq\_rel(A) = \frac{\sum_{1}^{n} oA}{n}\ , \qquad (3.1)$$

onde: oA = ocorrência de A na tabela de transações,

n = número de transações.

$$FREAB = FRAxFRB, \tag{3.2}$$

onde: FRA = freqüência relativa de A, calculada conforme (3.1) e

FRB = freqüência relativa de B, calculada de acordo com (3.3).

$$FRB = freq\_rel(B) = \frac{\sum_{1}^{n} oB}{n}, \tag{3.3}$$

onde: oB = ocorrência de B na tabela de transações.

$$FROAB = freq\_relO(A\hat{}B) = \frac{\sum_{1}^{n} (oA \hat{} oB)}{n}, \tag{3.4}$$

onde: oA^oB = ocorrência de A e B, simultaneamente.

3.1.2 - Fuzificação

No processo de fuzificação, FRA, FREAB e FROAB são transformadas em variáveis difusas, compostas por três conjuntos, representados pelos termos lingüísticos: “alta”, “moderada” e “baixa”. Devido ao fato de as três entradas terem as mesmas características, a fuzificação é feita de forma idêntica e os intervalos de funções de pertinência são ilustrados na Figura 3.2. A investigação que resultou na escolha dessas funções e intervalos encontra-se na Seção 4.2.

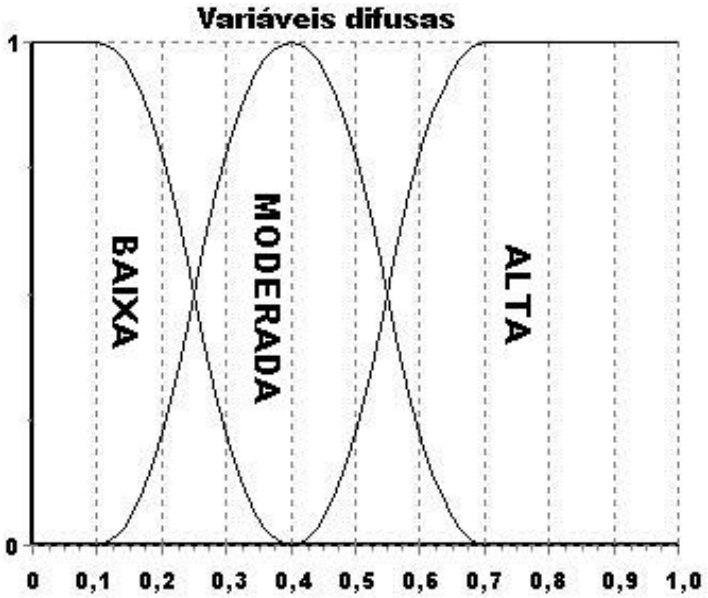


Figura 3.2: Representação dos conjuntos difusos para as três entradas numéricas.

Conforme Figura 3.2, o conjunto que representa o termo lingüístico "baixa" é limitado pela função definida por (2.26), com α=0,1 e β=0,3; "moderada", pela função definida por (2.25), com α=0,4 e β=0,3 ; e "alta", pela função definida por (2.27), com α=0,7 e β=0,3. O processo de fuzificação ocorre da seguinte maneira: para cada entrada é calculado um valor de pertinência μ, para cada conjunto; e os valores obtidos são armazenados para serem usados na etapa de propagação. A Tabela 3.1 mostra três exemplos de fuzificação.

Tabela 3.1: Exemplos de fuzificação dos valores numéricos

Entradas	***Valor Numérico***	μ_{Baixa}	$\mu_{Moderada}$	μ_{Alta}
FRA	0,780	0,000	0,000	1,000
	0,220	0,320	0,680	0,000
	0,170	0,755	0,245	0,000
FREAB	0,320	0,000	0,980	0,02
	0,060	1,000	0,000	0,000
	0,070	1,000	0,000	0,000
FROAB	0,120	0,980	0,002	0,000
	0,210	0,405	0,595	0,000
	0,140	0,920	0,080	0,000

Os termos apresentados na Tabela 3.1 têm os seguintes significados:

- FRA – Freqüência Relativa do antecedente da regra “se A então B”;
- FREAB – Freqüência Relativa Esperada de A e B;
- FROAB – Freqüência Relativa Obtida de A e B;
- Valor numérico – Valor de entrada de cada freqüência acima;
- μ_{Baixa} - Grau de pertinência do valor de entrada ao conjunto difuso cujo termo lingüístico é “Baixa”;
- $\mu_{Moderada}$ - Grau de pertinência do valor de entrada ao conjunto difuso cujo termo lingüístico é “Moderada”;
- μ_{Alta} - Grau de pertinência do valor de entrada ao conjunto difuso cujo termo lingüístico é “Alta”.

A Figura 3.3 ilustra um dos exemplos, onde: (a) representa a fuzificação do valor 0,780 da variável FRA; (b), do valor 0,320 da variável FREAB fuzificada; (c), do valor 0,120 da variável FROAB fuzificada.

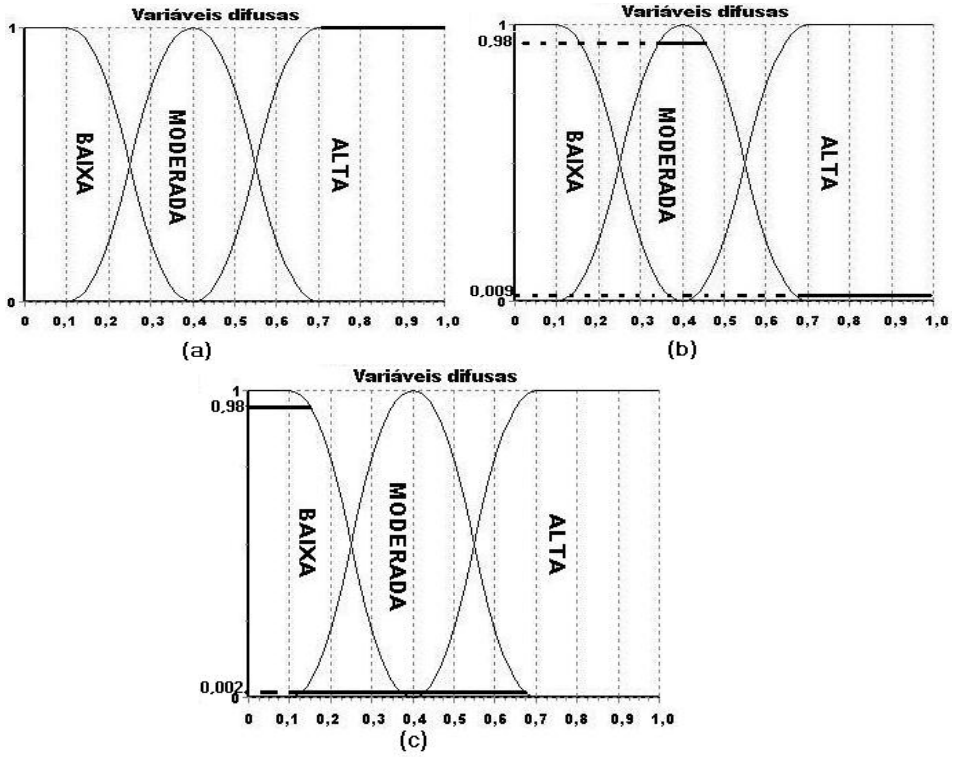


Figura 3.3: Exemplo de fuzificação das variáveis de entrada.

3.1.3 - Propagação

Nesta etapa ocorre o processo de inferência difusa, no qual as regras apresentadas a seguir são disparadas de acordo com os graus de pertinência obtidos no processo de fuzificação e com um método de composição. Ressalta-se que foi investigado qual o método de composição mais adequado, conforme mostrado na Seção 4.1.7.

Regras

As regras foram obtidas através de *heurística.* Primeiro, montou-se uma árvore de possibilidades para combinar as três entradas, resultando em 27 possíveis regras, visto que há três entradas e três conjuntos difusos para cada entrada, do qual se tem 3^3 possibilidades. Em seguida foi selecionado, por sorteio, um conjunto de dados que permitiram deduzir a saída para cada regra e descartar algumas possibilidades que não ocorrem na prática como, por exemplo, ter as duas primeiras entradas baixas e a última alta. As regras obtidas, apresentadas de forma resumida na Tabela 3.2, são as seguintes:

1. *Se **FRA** é BAIXA e **FREAB** é BAIXA e **FROAB** é BAIXA, então **MDCAR** é Atração Moderada.*
2. *Se **FRA** é BAIXA e **FREAB** é MODERADA e **FROAB** é BAIXA, então **MDCAR** é Repulsão Alta.*
3. *Se **FRA** é BAIXA e **FREAB** é MODERADA e **FROAB** é MODERADA, então **MDCAR** é Atração Moderada.*
4. *Se **FRA** é MODERADA e **FREAB** é BAIXA e **FROAB** é BAIXA, então **MDCAR** é Repulsão Baixa.*
5. *Se **FRA** é MODERADA e **FREAB** é BAIXA e **FROAB** é MODERADA, então **MDCAR** é Atração Alta.*
6. *Se **FRA** é MODERADA e **FREAB** é MODERADA e **FROAB** é BAIXA, então **MDCAR** é Repulsão Alta.*
7. *Se **FRA** é MODERADA e **FREAB** é MODERADA e **FROAB** é MODERADA, então **MDCAR** é Atração Baixa.*
8. *Se **FRA** é MODERADA e **FREAB** é ALTA e **FROAB** é BAIXA, então **MDCAR** é Repulsão Alta.*

9. *Se **FRA** é MODERADA e **FREAB** é ALTA e **FROAB** é MODERADA, então **MDCAR** é Repulsão Alta.*
10. *Se **FRA** é ALTA e **FREAB** é BAIXA e **FROAB** é BAIXA, então **MDCAR** é Repulsão Baixa.*
11. *Se **FRA** é ALTA e **FREAB** é BAIXA e **FROAB** é MODERADA, então **MDCAR** é Atração Moderada.*
12. *Se **FRA** é ALTA e **FREAB** é MODERADA e **FROAB** é BAIXA, então **MDCAR** é Repulsão Moderada.*
13. *Se **FRA** é ALTA e **FREAB** é MODERADA e **FROAB** é MODERADA, então **MDCAR** é Repulsão Baixa.*
14. *Se **FRA** é ALTA e **FREAB** é MODERADA e **FROAB** é ALTA, então **MDCAR** é Atração Alta.*
15. *Se **FRA** é ALTA e **FREAB** é ALTA e **FROAB** é BAIXA, então **MDCAR** é Repulsão Alta.*
16. *Se **FRA** é ALTA e **FREAB** é ALTA e **FROAB** é MODERADA, então **MDCAR** é Repulsão Moderada.*
17. *Se **FRA** é ALTA e **FREAB** é ALTA e **FROAB** é ALTA, então **MDCAR** é Atração Baixa.*

Tabela 3.2: Resumo das regras de inferência, difusas

Regra	FRA	FREAB	FROAB	MDCAR
1	BAIXA	BAIXA	BAIXA	AMODERADA
2	BAIXA	MODERADA	BAIXA	RALTA
3	BAIXA	MODERADA	MODERADA	AMODERADA
4	MODERADA	BAIXA	BAIXA	RBAIXA
5	MODERADA	BAIXA	MODERADA	AALTA
6	MODERADA	MODERADA	BAIXA	RALTA
7	MODERADA	MODERADA	MODERADA	ABAIXA
8	MODERADA	ALTA	BAIXA	RALTA
9	MODERADA	ALTA	MODERADA	RALTA
10	ALTA	BAIXA	BAIXA	ABAIXA
11	ALTA	BAIXA	MODERADA	ABAIXA
12	ALTA	MODERADA	BAIXA	RMODERADA
13	ALTA	MODERADA	MODERADA	RBAIXA
14	ALTA	MODERADA	ALTA	AMODERADA
15	ALTA	ALTA	BAIXA	RALTA
16	ALTA	ALTA	MODERADA	RMODERADA
17	ALTA	ALTA	ALTA	ABAIXA

Os termos apresentados na Tabela 3.2 têm os seguintes significados:

- FRA – Freqüência Relativa do antecedente da regra “se A então B”;
- FREAB – Freqüência Relativa Esperada de A e B;
- FROAB – Freqüência Relativa Obtida de A e B;
- MDCAR – Método Difuso para Cálculo de Atração/Repulsão.

3.1. 4 - Classificação

Nesta etapa, as regras disparadas na etapa anterior são compostas por “Max-Produto” por ser a forma de composição que apresentou melhores resultados – conforme discutido na seção 4.2.

A variável de saída, chamada MDCAR e ilustrada na Figura 3.4, é composta por 6 conjuntos difusos representados pelos termos lingüísticos: RALTA (Repulsão Alta) limitado pela curva definida por (2.23); RMOD (Repulsão Moderada) limitado pela curva definida por (2.22); RBAIXA (Repulsão Baixa) limitado pela curva definida por (2.24); ABAIXA (Atração Baixa) limitado pela curva definida por (2.22); AMOD (Atração Moderada), limitado pela curva definida por (2.22); e AALTA (Atração Alta) limitado pela curva definida por (2.24).

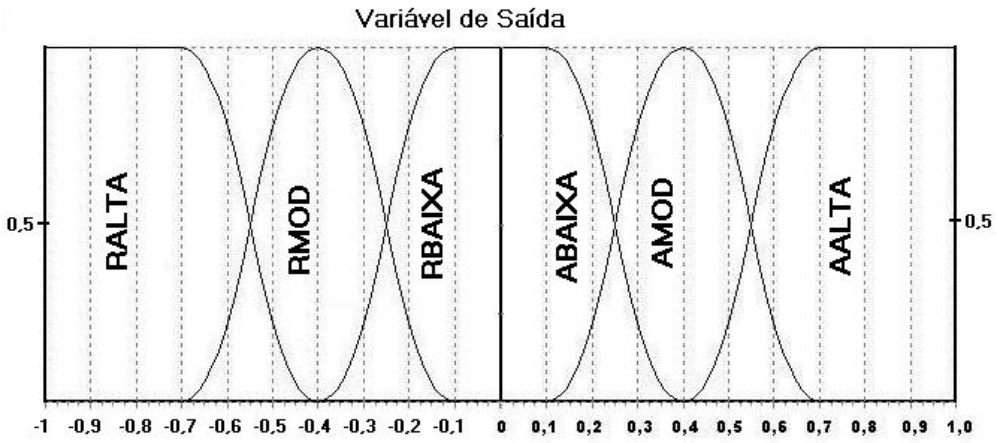


Figura 3.4: Representação dos conjuntos difusos para a variável de saída.

Os termos apresentados na Figura 3.4 têm os seguintes significados:

- RALTA – Repulsão Alta entre os itens pesquisados;

- RMOD – Repulsão Moderada entre os itens pesquisados;
- RBAIXA – Repulsão Baixa entre os itens pesquisados;
- ABAIXA – Atração Baixa entre os itens pesquisados;
- AMOD – Atração Moderada entre os itens pesquisados;
- AALTA – Atração Alta entre os itens pesquisados.

Esse processo de classificação ocorre da seguinte forma:

1º) avalia-se todas as regras, calculando-se a saída de cada uma, através do produto entre os graus de pertinência μ_s calculados na etapa de fuzificação. A Tabela 3.3 mostra as saídas das regras para as entradas da Tabela 3.1;

2º) os valores de saída de cada regra são comparados, a que tiver o maior valor é a vencedora, sendo, portanto, a sua saída o resultado que será apresentado em termos lingüísticos. Por exemplo, a "saída(1)" da Tabela 3.3 mostra que a regra vencedora é a 12, com isso a saída em termos lingüísticos "Repulsão Moderada"; na "saída(2)", o resultado é "Atração Alta"; e na "saída(3)", "Atração Moderada". Estes termos lingüísticos representam os conjuntos da variável de saída, descrita a seguir.

Tabela 3.3: Valores de saída para cada regra

Regra	Saída(1)	Saída(2)	Saída(3)
1	0,000	0,130	0,695
2	0,000	0,000	0,000
3	0,000	0,000	0,000
4	0,000	0,280	0,225
5	0,000	0,400	0,020
6	0,000	0,000	0,000
7	0,000	0,000	0,000
8	0,000	0,000	0,000
9	0,000	0,000	0,000
10	0,000	0,000	0,000
11	0,000	0,000	0,000
12	0,960	0,000	0,000
13	0,002	0,000	0,000
14	0,000	0,000	0,000
15	0,009	0,000	0,000
16	0,000	0,000	0,000
17	0,000	0,000	0,000

Os termos apresentados na Tabela 3.3 têm os seguintes significados:

- Saída(1) – Saída obtida a partir das entradas FRA = 0,780, FREAB = 0,320 e FROAB = 0,120 mostradas na Tabela 3.1;
- Saída(2) – Saída obtida a partir das entradas FRA = 0,220, FREAB = 0,060 e FROAB = 0,210 mostradas na Tabela 3.1;
- Saída(3) – Saída obtida a partir das entradas FRA = 0,170, FREAB = 0,070 e FROAB = 0,140 mostradas na Tabela 3.1.

A Figura 3.5 ilustra o processo de classificação, usando como exemplo a "saída(2)" da Tabela 3.3, para a qual são disparadas as regras 1, 4 e 5. Conforme pode ser observado, a regra 5 é a vencedora, portanto a saída é atração alta.

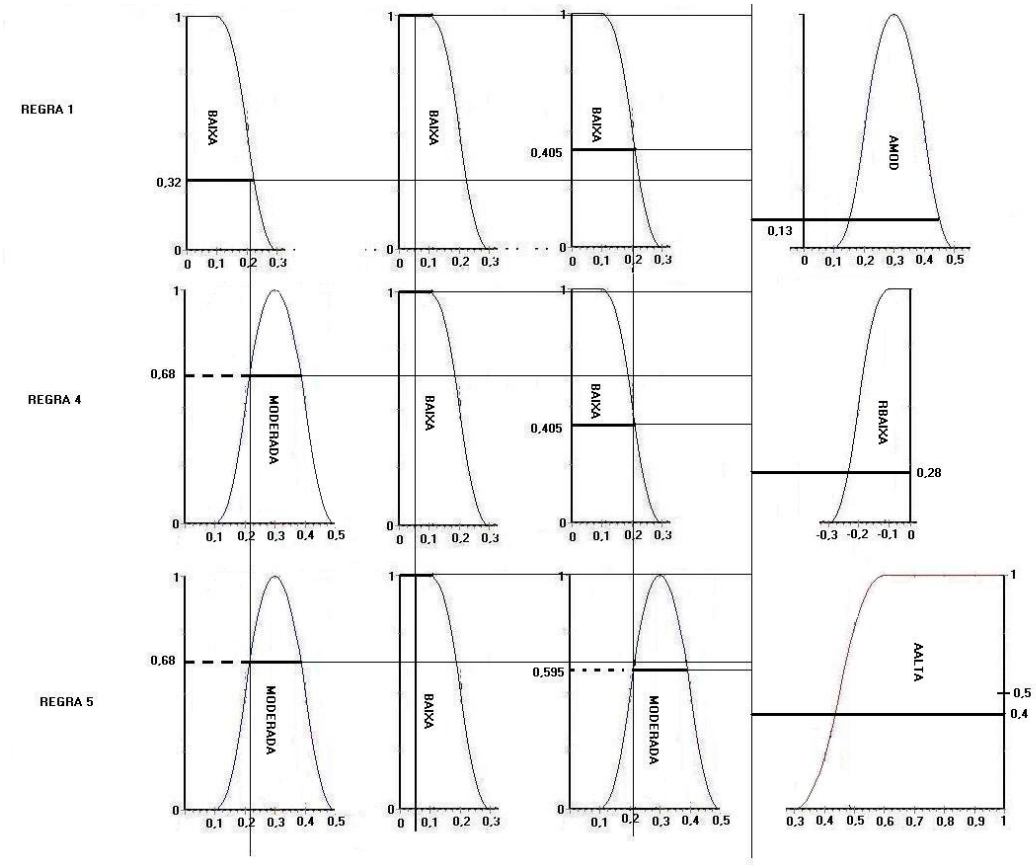


Figura 3.5: Ilustração gráfica da etapa de classificação.

3.2 – Utilização do MDCAR

O MDCAR pode ser usado para fazer consulta personalizada a uma base de dados, por exemplo, pode-se criar uma interface que permita ao usuário entrar com consulta da seguinte forma: “o item A tem forte atração pelo item B?” ou “o item C é atraído pelos itens A e B?” ou, então, “quais são os itens que têm maior atração pelo item W?”, dentre outras.

A título de exemplificação do uso do MDCAR, seja a Tabela 3.4, onde A, B e C representam os produtos comercializados, seja a transação que representa cada cesta de compra do estabelecimento. Suponha que o usuário faça essa consulta: “o produto B tem forte atração pelo produto A?”, a resposta é sim, porque a resposta fornecida pelo MDCAR é AALTA, que significa que A atrai B fortemente, ou seja, que B tem forte atração por A. Pode-se fornecer mais informações ao usuário como, por exemplo, o grau de pertinência ao conjunto AALTA ou, talvez, fornecer o valor da saída *numérico* também. Isso depende de cada analista que for usar o método em seu sistema.

Tabela 3.4: Tabela exemplo de transações

Transação	**A**	**B**	**C**
1	1	1	0
2	0	0	1
3	0	0	1
4	1	1	1
5	1	1	0
6	0	0	0
7	1	1	1
8	0	0	1
9	0	0	0
10	0	0	1

A interface que tornaria possível a consulta mencionada no parágrafo anterior deve disparar uma função que busque, na matriz de co-ocorrência (exemplificada pela Tabela 3.4), os valores de FRA, FREAB e FROAB e os forneça para o método MDCAR a fim de disparar todo o processo descrito na Seção 3.1.

Para o desenvolvimento do MDCAR foi necessário testar vários modelos difusos, várias combinações de funções de pertinências e intervalos, até conseguir decidir qual é

a melhor combinação para o método proposto. Para isso foram necessárias centenas de ensaios. Na próxima seção, são apresentados todos os passos seguidos para se chegar ao modelo apresentado aqui.

Capítulo 4 – Ensaios e Resultados

Neste capítulo, são apresentados os ensaios realizados com vários modelos de composição difusa e combinações de funções de pertinência e os correspondentes intervalos. É apresentada também uma análise dos resultados obtidos, procurando mostrar o modelo mais adequado à resolução do problema em questão. Para realizar estes ensaios, o método MDCAR foi implementado em DELPHI^{TM} Versão 7.0.

Para realizar uma avaliação do método MDCAR, realizou-se uma comparação com a medida de *lift* de Groth e com a obtida pelo MDCAR. Assim foi necessário obter a estimativa da medida *lift* por meio de uma etapa de desfuzificação conforme figura 4.1. Esta etapa de desfuzificação permite obter um valor numérico para a medida *lift*.

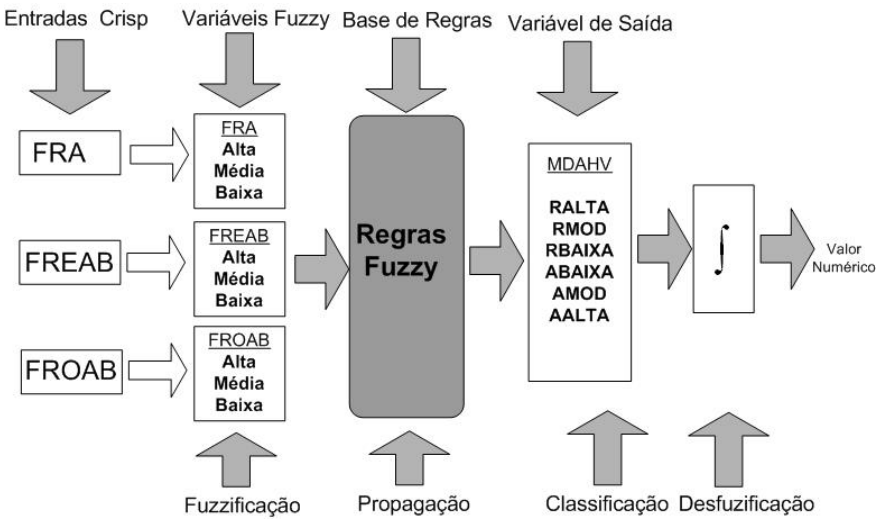


Figura 4.1: Modelo usado para os testes.

4.1 – Ensaios Realizados

Para se chegar a um conjunto de dados para desenvolvimento e testes do método proposto, foram necessárias várias etapas, como são apresentadas nas próximas seções.

4.1.1 – Aquisição dos Dados

Um dos grandes problemas para se desenvolver pesquisa para aplicação comercial está na aquisição de dados, porque as empresas têm receio de fornecê-los, por suspeitarem que isso possa ser prejudicial ao seu negócio.

Outro grande problema é o formato dos dados, devido aos SGDBs usados pelas empresas, nesse caso, *Oracle*, *Interbase* e *SQL Server*. Para resolver tal problema, foi necessário desenvolver um programa para se comunicar com os três formatos ao mesmo tempo e colocá-los em um único formato, para facilitar o trabalho de implementação do método MDCAR.

Os dados usados na presente pesquisa foram obtidos do histórico de vendas, de cinco anos, de uma rede de supermercados, de uma loja de departamentos e de uma rede de livrarias.

A massa de dados adquirida contém milhões de transações (linhas) e milhares de itens (colunas), mas nem todos foram usados. Para agilizar o processo, foram selecionadas amostras de cada base, conforme é apresentado na próxima seção.

4.1.2 – Seleção dos dados

Foi escolhida uma amostra, por sorteio, entre itens e transações cujo tamanho se encontra na Tabela 4.1. O porquê da grande diferença entre quantidade de itens e de transações reside no fato de existirem mais transações do que itens no histórico de compras. Além disso, o esforço computacional para trabalhar número de itens elevados é muito maior do que é para trabalhar com elevado número de transações, como mostrado na Seção 4.1.4.

Tabela 4.1: Tamanho das amostras

	Tamanho
Itens	1.100
Transações	600.000

4.1.3 – Purificação dos Dados

Nesta etapa, os dados selecionados foram inspecionados com ajuda de recursos computacionais. E as transações que continham dados perdidos ou incompletos foram eliminadas e as que continham dados inconsistentes foram corrigidas, utilizando como base as transações similares e que continham dados consistentes. Para cada transação eliminada, outra foi escolhida, por sorteio, para substituí-la.

4.1.4 – Transformação dos Dados

Após as etapas anteriores, os dados purificados foram dispostos em tabelas, cada base em uma tabela distinta, como exemplifica a Tabela 4.2.

Tabela 4.2: Exemplo de histórico de compra após pré-processamento

Cliente	**Compra**			
001	Cinto	Carteira	Sapato	...
002	Calça	Camisa	Sapato	...
003	Relógio	Meias	Camisa	...
...	...	...	...	...
600.000	Calça	Relógio	...	...

Como o objetivo desta etapa é transformar os dados para se ajustarem ao algoritmo de mineração a ser aplicado, foi necessário rearranjar as entradas da tabela 4.2, de forma que se tenham apenas zeros e uns, que são os dados adequados a este tipo de problema. A Tabela 4.3 mostra como os dados mostrados na Tabela 4.2 (conhecida por matriz de co-ocorrência) ficaram após a transformação. Os detalhes, dessa transformação, podem ser encontrados na Seção 2.4.1.

Tabela 4.3: Matriz de co-ocorrência para os produtos da Tabela 4.2

Cliente	Cinto	Carteira	Sapato	Calça	Camisa	Relógio	Meias	...
001	1	1	1	0	0	0	0	...
002	0	0	1	1	1	0	0	...
003	0	0	0	0	1	1	1	...
...	...	...	...	...	...	...	...	...
600.000	0	0	0	1	0	1	0	...

Após a transformação dos dados onde foi obtida a matriz de co-ocorrência, conforme Tabela 4.3, inicia-se a próxima etapa, que envolve a descoberta do conhecimento, que nesse caso é a procura por valores relativos às três entradas para o método proposto, que são: Freqüência Relativa de A, Freqüência Relativa Esperada de A e B e Freqüência Relativa Obtida de A e B.

4.1.5 – Obtenção dos Dados de Entrada para o MDCAR

A partir desta etapa, não importa a origem dos dados, o que importa é obter um conjunto de valores para as três entradas. Para isso, pensou-se em calcular os valores de todas as associações possíveis para cada amostra, o que se mostrou inviável devido ao número elevado de combinações possíveis, o que pode ser observado por (4.3). Para obter (4.3), usou-se o princípio de indução matemática.

$$A = n! + \sum_{j=2}^{n-1} N_j \,, \tag{4.3}$$

onde: A = número de associações possíveis

n = número de atributos da matriz de co-ocorrência

os N_{js} são calculados por (4.4).

$$N_j = 2(n-j)\sum_{i=j-1}^{n-1}(n-i) \;;\; 2 \le j \le n-1, \tag{4.4}$$

onde: N = nível de associação, explicado a seguir.

Para explicar os níveis de associação calculados por (4.4), seguem os exemplos abaixo:

Exemplo 1: Suponha que a matriz de co-ocorrência tenha dois atributos (A e B), tem-se apenas associações do primeiro nível, que são: $A \rightarrow B$ e $B \rightarrow A$, ou seja, duas associações possíveis (2! = 2).

Exemplo 2: Suponha agora 3 atributos (A, B e C), tem-se associações do primeiro e segundo níveis, que são: $A \rightarrow B$, $B \rightarrow A$, $A \rightarrow C$, $C \rightarrow A$, $B \rightarrow C$, $C \rightarrow B$ (primeiro nível); $AB \rightarrow C$, $AC \rightarrow B$, $CB \rightarrow A$, $C \rightarrow AB$, $B \rightarrow AC$ e $A \rightarrow CB$ (segundo nível), ou seja, 12 associações possíveis, 3! + N_2 = 12, onde $N_2 = 2(3-2)[(3-1)+(3-2)]) = 6$.

À medida que se aumenta o número de atributos, o número de associações possíveis aumenta exponencialmente, conforme ilustrado pela Figura 4.2; portanto para 1.100 atributos (itens), é inviável trabalhar com todas as associações possíveis. A solução encontrada foi retirar uma amostra (de 600.000 associações, com o mesmo número de transações de cada base de dados) de cada conjunto de dados.

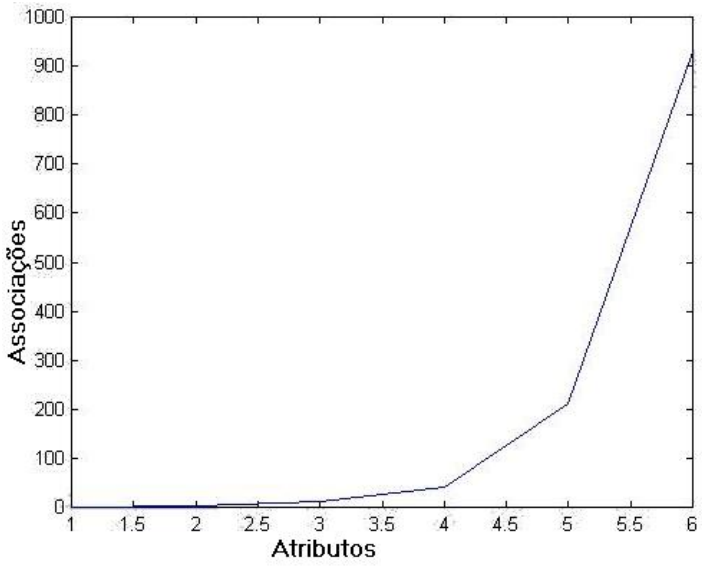


Figura 4.2: Gráfico - atributos X associações.

Das 600.000 associações possíveis, retiradas de cada conjunto de dados (no total foram 1.800.000), foram calculados o FRA, FREAB e FROAB, que são as entradas

usadas para o desenvolvimento do MDCAR. Desse total, foram retirados 10% para servir como base de teste, ficando, assim, a base de trabalho com 1.620.000 e a base de teste com 180.000. Desses totais, foram retiradas associações com freqüências relativas, para qualquer uma das três variáveis, com valores inferiores a 5%, sobrando, dessa forma, a base de trabalho com 1.242.507 linhas e a base de teste com 124.200 linhas. Essas bases de dados e as demais obtidas nos ensaios se encontram no CD, em anexo, e as explicações sobre o conteúdo se encontram no Apêndice A.

A Tabela 4.4 mostra um estrato da base de dados de trabalho. Os valores apresentados na tabela são as freqüências relativas de cada entrada.

Tabela 4.4: Exemplos de dados da base de trabalho

FRA	FREAB	FROAB
0,47	0,06	0,09
0,70	0,20	0,14
0,56	0,28	0,37
...	...	...

4.1.6 –Intervalos

Os intervalos foram determinados por meio de heurísticas. Foi fixado um valor para cada entrada e foram testados vários intervalos e as saídas foram observadas e analisadas, a fim de verificar se havia coerência ou não com o que se esperava obter. Os intervalos para cada combinação de funções de pertinência são mostrados na Tabela B.1, que se encontra no Apêndice B. O modelo de composição usado foi o de *Max-Produto,* por ser o que se mostrou mais adequado durante testes preliminares realizados.

Os intervalos, valores entre 0 e 1, obtidos para cada função são mostrados na Tabela 4.5. Todas essas funções foram usadas para determinação do modelo difuso de composição ideal para compor o MDCAR, como é mostrado na próxima seção.

Tabela 4.5: Parâmetros das funções de pertinência usados para os ensaios

	Parâmetros					
Função de pertinência	**a**	**b**	**c**	**d**	α	β
L	0,00	0,10	0,40	-	-	-
Gama	0,40	0,70	1,00	-	-	-
Triangular	0,10	0,40	0,70	-	-	-
Trapezoidal	0,10	0,35	0,45	0,70	-	-
Gaussiana	-	-	-	-	0,40	0,30
Sigmoidal	-	-	-	-	0,70	0,30
Z	-	-	-	-	0,10	0,30
Sino	0,10	0,35	0,45	0,70	-	-

4.1.7 – Funções de Pertinência

As funções, para cada conjunto (Baixa, Moderada e Alta) – que representam as variáveis lingüísticas para as entradas do sistema (FRA, FREAB e FROAB), foram escolhidas de acordo com suas características. Para o conjunto difuso “baixa”, as funções que mais se ajustaram foram L e Z (ilustradas pelas figuras B.1 e B.6, respectivamente); para o conjunto difuso “alta”, as funções que mais se ajustaram foram gama e sigmoidal (ilustradas pelas figura B.2 e B.7); as demais, triangular (Figura B.3) , trapezoidal (Figura B.4), Pi (Figura B.5) e sino (Figura B.8), são mais indicadas para o conjunto difuso “moderada”.

Para se chegar a uma combinação de funções ideal, foram testadas todas as combinações possíveis, 16 no total, das funções mostradas na Tabela 4.5. A Tabela 4.6 mostra as combinações de funções, cujos gráficos se encontram no apêndice D, usadas nesta pesquisa.

Para cada combinação de funções, foram testados vários modelos de composição e desfuzificação, como é apresentado na próxima seção.

Tabela 4.6: Combinações entre as principais funções de pertinência

	Conjunto		
Combinação	**Baixa**	**Moderada**	**Alta**
1	L	Triangular	Gama
2			Sigmoidal
3		Trapezoidal	Gama
4			Sigmoidal
5		PI	Gama
6			Sigmoidal
7		Sino	Gama
8			Sigmoidal
9	Z	Triangular	Gama
10			Sigmoidal
11		Trapezoidal	Gama
12			Sigmoidal
13		PI	Gama
14			Sigmoidal
15		Sino	Gama
16			Simoidal

4.1.8 – Modelos Difusos

Os modelos difusos de composição testados foram: o modelo de classificação; o modelo de *Mamdani;* o modelo de *Takagi-Sugeno;* e o modelo de *Tsukamoto;* com composição de regras do tipo *mínimo* e *produto* (produto algébrico, conforme (2.20)) para cada um. Para cada uma das dezesseis combinações da Tabela 4.6, foram testados todos esses modelos duas vezes, uma com composição *mínimo* e outra com *produto*, dando um total de 128 repetições dos testes com a mesma base, base de testes com 124.200 linhas. Os resultados foram gravados em bases de dados que se encontram no CD, em anexo.

4.1.8.1 - Classificação

No modelo de classificação, como ilustrado pela Figura 2.10 – Seção 2.5.2.2, não há a etapa de desfuzificação. Calcula-se a t-norma (interseção padrão) – nesta pesquisa

testou-se a interseção produto algébrico também, para cada regra disparada – em seguida calcula-se a t-conorma, ou seja, busca-se o máximo entre estes mínimos e a regra vencedora, a qual apresentou o maior valor para a t-conorma e determina a saída do sistema. A seguir é apresentado um exemplo do funcionamento do método, usando o modelo de classificação.

Suponha as entradas 0,52, 0,36 e 0,28 para FRA, FREAB e FROAB, respectivamente. Usando a combinação de funções 14 (Z, Pi, Sigmoidal), mostrado na Tabela 4.6, com os parâmetros apresentados na Tabela 4.5, obtêm-se os graus de pertinência para cada conjunto apresentados na Tabela 4.7.

Tabela 4.7: Graus de pertinência para cada entrada

Entradas	***Valor Numérico***	μ_{Baixa}	$\mu_{Moderada}$	μ_{Alta}
FRA	0,52	0,00	0,00	0,86
FREAB	0,36	0,00	0,82	0,08
FROAB	0,28	0,02	0,98	0,002

Os termos apresentados na Tabela 4.7 têm os seguintes significados:

- FRA – Freqüência Relativa do antecedente da regra “se A então B”;
- FREAB – Freqüência Relativa Esperada de A e B;
- FROAB – Freqüência Relativa Obtida de A e B;
- Valor numérico – Valor de entrada de cada freqüência acima;
- μ_{Baixa} - Grau de pertinência do valor de entrada ao conjunto difuso cujo termo lingüístico é “Baixa”;
- $\mu_{Moderada}$ - Grau de pertinência do valor de entrada ao conjunto difuso cujo termo lingüístico é “Moderada”;
- μ_{Alta} - Grau de pertinência do valor de entrada ao conjunto difuso cujo termo lingüístico é “Alta”.

Usando mínimo ou produto (para comparação entre ambos), as regras disparadas foram 12, 13, 15 e 16, sendo que a regra 13 possui o maior grau de pertinência (como mostra a Tabela 4.8); portanto, é a regra vencedora e resultado o “repulsão baixa”, conforme conjunto de regras mostrado na Seção 3.1.3.

Tabela 4.8: Regras disparadas

Composição	Regra	Grau
Mínimo	12	0,02
	13	0,82
	14	0,002
	15	0,02
	16	0,08
Produto	12	0,014
	13	0,69
	14	0,001
	15	0,0013
	16	0,067

A partir desta saída do método MDCAR, a resposta dada ao usuário poderia ser na forma: “O produto A repele fracamente o produto B” . E fica a cargo da criatividade de cada analista.

4.1.8.2 - Mamdani

O modelo de *Mamdani* difere do modelo de classificação (Figura 2.16) pelo fato de permitir a desfuzificação, que é feita como ilustra a pela Figura 2.6, mostrada na Seção 2.5.2.2.

Usando as mesmas entradas da seção anterior, a classificação é “repulsão baixa”, para os dois casos – mínimo e produto. Após a desfuzificação, as saídas numéricas obtidas foram as seguintes: usando o padrão do modelo de *Mamdani*, teve-se como resultado o valor de -0,123; mudando o padrão, em vez de mínimo, usando o produto, o resultado obtido foi de -0,104. O valor calculado utilizando *lift* de Groth (2.5) foi -0,154.

4.1.8.3 - Tsukamoto

O modelo de *Tsukamoto* difere dos modelos anteriores, por exigir que as funções de pertinência para cada conjunto de saída sejam monotônicas. Para isso, as funções e intervalos de saída foram adaptados, conforme ilustra a Figura 4.3. Para os demais modelos, os conjuntos da variável de saída são os mesmos conforme ilustra a Figura 3.5. O modelo de *Tsukamoto* exige que as funções associadas aos conjuntos de saída, sejam monotônicas, ou seja, estritamente crescente ou estritamente decrescente. Devido a isso, buscou-se transformar as funções associadas aos conjuntos RMOD e AMOD em funções monotônicas, como ilustra a Figura 4.3. O uso de funções lineares (para este caso), em vez de não-lineares, deve-se ao fato da complexidade matemática, nesse caso, para se calcular as inversas das funções de pertinência.

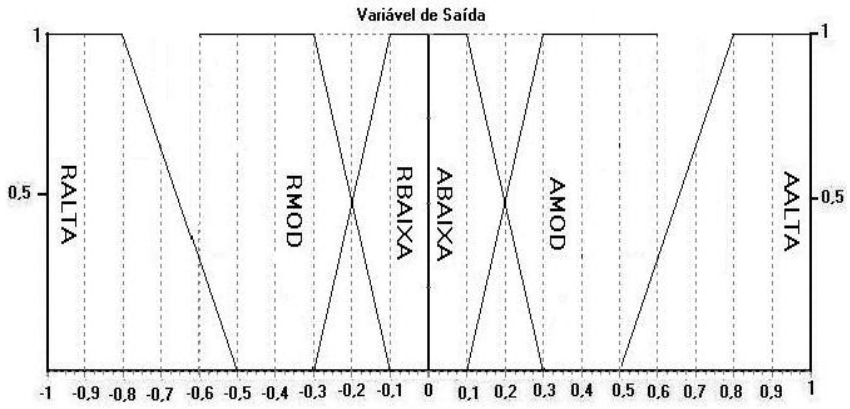


Figura 4.3: Variável de saída para o modelo de Tsukamoto.

A saída numérica final , usando o modelo de *Tsukamoto,* é calculada por (4.5), e o resultado, usando mínimo, é -0,099 e para produto, -0,152.

$$S_F = \frac{\sum_{i=1}^{n} s_i * w_i}{\sum_{i=1}^{n} w_i}, \qquad (4.5)$$

onde: n = nº de regras disparadas,

w_i = resultado da operação composição, mínimo ou produto para cada regra disparada,

S_F = saída *numérico* final,

s_i é calculado pela inversa da função monotônica de cada conjunto.

A Tabela 4.9 mostra as funções e suas inversas para cada conseqüente da regra, que são os conjuntos difusos de saída.

Tabela 4.9: Funções de pertinência e suas inversas

Conjunto	Função	Inversa
RALTA	$\mu(x) = \frac{-\frac{3}{10} - x}{\frac{7}{10}}$	$\mu^{-1}(x) = \frac{-7x-3}{10}$
RMODERADA	$\mu(x) = \frac{-\frac{1}{10} - x}{\frac{4}{10}}$	$\mu^{-1}(x) = \frac{-4x-1}{10}$
RBAIXA	$\mu(x) = \frac{x + \frac{5}{10}}{\frac{5}{10}}$	$\mu^{-1}(x) = \frac{5x-5}{10}$
ABAIXA	$\mu(x) = \frac{\frac{5}{10} - x}{\frac{5}{10}}$	$\mu^{-1}(x) = \frac{5-5x}{10}$
AMODERADA	$\mu(x) = \frac{x - \frac{1}{10}}{\frac{4}{10}}$	$\mu^{-1}(x) = \frac{4x+1}{10}$
AALTA	$\mu(x) = \frac{x - \frac{3}{10}}{\frac{7}{10}}$	$\mu^{-1}(x) = \frac{7x+3}{10}$

Os termos apresentados na Tabela 4.9 têm os seguintes significados:

- RALTA – Repulsão Alta entre os itens pesquisados;
- RMOD – Repulsão Moderada entre os itens pesquisados;
- RBAIXA – Repulsão Baixa entre os itens pesquisados;
- ABAIXA – Atração Baixa entre os itens pesquisados;
- AMOD – Atração Moderada entre os itens pesquisados;

- AALTA – Atração Alta entre os itens pesquisados.

As funções mostradas na Tabela 4.9 foram obtidas através de combinações lineares que mapeiam as entradas para cada saída.

4.1.8.4 – Takagi-Sugeno

Para o modelo de *Takagi-Sugeno*, usando as mesmas entradas da Tabela 3.7, teve-se como resultado a "repulsão baixa" e a saída numérica para mínimo igual a -0,135 e, para produto, -0,14. O processo de fuzificação e classificação foram apresentados na Seção 3.1 e o processo de desfuzificação é apresentado a seguir.

Desfuzificação

Nesta etapa, a saída difusa, isto é, na forma de termos lingüísticos, é transformada em um valor numérico. Esta é realizada, seguindo o seguinte procedimento:

1. calcula-se as saídas parciais de cada regra através da combinação linear das entradas, conforme (3.5);

$$S_i = xp_i + yq_i + zr_i + s_i, \tag{4.6}$$

onde: S_i = saída parcial para cada regra

i é o número da regra (Regra 1, Regra 2, ...)

x = FRA

y = FREAB

z = FROAB

p,q,r e s são coeficientes lineares das entradas para cada regra, obtidos através de interpolação polinomial que se encontram na Tabela 4.10.

Tabela 4.10: Coeficientes lineares para cada regra

Regra	P	q	r	S
1	-0,820	-0,820	1,300	0,050
2	0,675	-0,880	0,675	-0,050
3	-0,800	-2,110	-1,890	0,102
4	0,170	-3,170	3,080	-0,040
5	-0,860	-4,100	2,610	0,560
6	1,140	-0,830	-2,330	0,570
7	0,221	-0,370	0,860	-0,240
8	2,670	-2,830	3,000	-0,960
9	3,680	-2,500	0,270	-0,690
10	-0,040	-0,950	1,550	-0,030
11	-1,690	0,390	1,670	0,810
12	0,420	-1,320	1,080	-0,300
13	0,120	-1,580	1,580	-0,070
14	-0,110	-0,210	0,176	0,079
15	1,590	-1,940	2,460	-0,860
16	0,300	-1,230	0,800	-0,080
17	0,150	-1,500	1,360	-0,010

2. calcula-se a saída final, que é a saída numérica desejada, de acordo com (3.6).

$$R_f = \frac{\sum_{i=1}^{17} w_i s_i}{\sum_{i=1}^{17} w_i}, \tag{4.7}$$

onde: R_f é a saída numérica final do método,

s_i = saídas parciais de cada regra,

w_i = pesos, valores obtidos através da operação *mínimo* ou *produto*, para cada regra mostrada na Tabela 3.2.

Suponha que se deseja calcular a saída numérica para as seguintes entradas 0,78, 0,32 e 0,12 para FRA, FREAB e FROAB, respectivamente. Os w_{is} estão determinados conforme mostra a Tabela 3.2, "saída(1)". Os s_{is} são calculados conforme (4.6) e são mostrados na Tabela 4.11. Com isso a saída final, calculada conforme (4.7), é igual a **-0,2623**, enquanto o *lift* calculado por (2.5) é igual a **-0,2564**. Isto mostra que o valor fornecido pelo método MDCAR está bem próximo da saída fornecida pelo *lift* de Groth.

Tabela 4.11: Saídas parciais para cada regra

Regra	Saída Parcial
1	-0,6960
2	0,2759
3	-1,4240
4	-0,5522
5	-1,1096
6	0,9140
7	-0,0828
8	0,5770
9	1,4128
10	-0,1792
11	-0,1830
12	-0,2652
13	-0,2924
14	-0,0529
15	0,0546
16	-0,1436
17	-0,2098

Apesar de a Tabela 4.11 mostrar as saídas parciais de todas as regras, não há necessidade do cálculo de todas, apenas das que foram disparadas. No exemplo acima seria necessário calcular só as saídas das regras 12, 13 e 15 que foram as regras disparadas. Embora isso não vá influenciar no resultado, visto que os pesos das regras não disparadas estão zerados.

Na próxima seção, são discutidos e analisados os resultados obtidos para cada modelo e para combinações de funções apresentadas aqui.

4.2 – Resumo do Método MDCAR Usado para os Testes

Seja o vetor de entrada V_E = (FRA, FREAB, FROAB).

1. *As entradas são transformadas em variáveis difusas, compostas por três conjuntos, representados pelos termos lingüísticos: “alta”, “moderada” e “baixa”.*

2. Calcula-se o grau de pertinência – aos conjuntos "alta", "baixa" e "moderada" – de cada valor de entrada.

3. Os valores da etapa anterior são propagados. Nessa etapa ocorre o processo de inferência difusa, em que as regras são disparadas.

4. As regras disparadas na etapa anterior são compostas por "Max-Produto", conforme os passos a seguir:

1º) avalia-se todas as regras, calculando-se a saída de cada uma, através do produto entre os valores calculados na etapa 2.

2º) compara-se os valores de saída de cada regra, a que apresentar o maior valor é a vencedora, sendo portanto sua saída o resultado que será apresentado em linguagem natural ao usuário. Por exemplo, se o conseqüente da regra vencedora for "Repulsão Alta", a resposta ao usuário é "o produto A repele fortemente o produto B".

5. A saída difusa é transformada em um valor numérico, usando o modelo de Takagi-Sugeno, conforme os passos a seguir:

1º) calcula-se as saídas parciais de cada regra através da combinação linear das entradas, conforme (4.5);

2º) calcula-se a saída final, que é a saída numérico desejada, de acordo com (4.6).

6. O resultado em termos lingüísticos e o valor numérico de saída são apresentados ao usuário.

4.3 - Resultados

Os resultados foram obtidos a partir de testes realizados, usando as combinações de funções mostradas na Tabela 4.6. Para a etapa de classificação do método proposto, foram testadas as composições *mínimo* e *produto*. Já, para a etapa de desfuzificação,

foram testados os modelos de *Mamdani*, *Takagi-Sugeno* e *Tsukamoto*, sendo que cada modelo foi testado com ambos os tipos de composição.

4.3.1 – Resultados Obtidos na Etapa de Classificação

Na etapa de classificação, foram observadas as saídas do sistema e comparadas com o valor numérico fornecido pelo MDCAR usado para teste e verificado se a saída é coerente ou não. Por exemplo, a Tabela 4.12 mostra que nas cinco primeiras linhas as saídas são coerentes porque a primeira linha mostra que o valor numérico de saída é igual a 0,03, o que pode ser considerado como baixa atração ou mesmo ausência de atração ou repulsão, pois está próximo de zero. O mesmo acontece com as linhas dois e quatro que, apesar de o primeiro caso sugerir baixa repulsão e o valor ser positivo, pode-se considerar como coerente a saída, porque o valor numérico está próximo de zero. Pode-se fazer a mesma análise para o segundo caso.

Tabela 4.12: Exemplos da etapa de classificação

Entradas			**Saídas**	
FRA	**FREAB**	**FROAB**	***Valor Numérico***	**MDCAR**
0,87	0,70	0,73	0,03	ABAIXA
0,57	0,10	0,11	0,01	RBAIXA
0,78	0,31	0,16	-0,20	RMODERA
0,50	0,455	0,450	-0,01	ABAIXA
0,67	0,55	0,18	-0,55	RALTA
0,74	0,40	0,53	0,17	AALTA

Os termos apresentados na Tabela 4.12 têm os seguintes significados:

- FRA – Freqüência Relativa do antecedente da regra “se A então B”;
- FREAB – Freqüência Relativa Esperada de A e B;
- FROAB – Freqüência Relativa Obtida de A e B;
- Valor numérico – Valor de entrada de cada freqüência acima;
- MDCAR – Método Difuso para Cálculo de Atração/Repulsão.

As saídas das linhas três e cinco também podem ser consideradas coerentes, pois esses valores pertencem aos conjuntos “repulsão moderada” e “repulsão alta”, respectivamente, com certo grau de pertinência. Já a saída da linha seis não pode ser considerada coerente, pois o valor 0,17 não pertence ao conjunto “atração alta”, porém isso se explica pelo fato da imprecisão contida na matriz de co-ocorrência. Pelo exemplo da Tabela 4.12, pode-se concluir que 83,3% das saídas são coerentes. Baseado nesse tipo de análise, chegou-se aos resultados descritos a seguir.

A Tabela 4.13 mostra os resultados de cada combinação de funções apresentados na Tabela 4.6 para os conjuntos de entrada. Como pode ser observado, não há grande diferença entre os dois modelos de composição.

Conforme mostra a Tabela 4.13 e ilustra a Figura 4.4, a melhor combinação foi a dois, composição *mínimo*, e a pior foi a onze, composição *produto*. Pode-se observar também que dependendo da combinação escolhida, o resultado foi o mesmo, não importando, portando, o tipo composição. O que importa, nesse caso, é o tipo de combinação de funções.

Tabela 4.13: Resultado dos testes para a etapa de classificação

Combinação	Mínimo	Produto
1	0,801	0,801
2	0,849	0,801
3	0,794	0,794
4	0,793	0,793
5	0,793	0,793
6	0,794	0,794
7	0,788	0,785
8	0,789	0,786
9	0,801	0,801
10	0,801	0,801
11	0,796	0,702
12	0,795	0,795
13	0,793	0,793
14	0,794	0,794
15	0,790	0,787
16	0,790	0,788

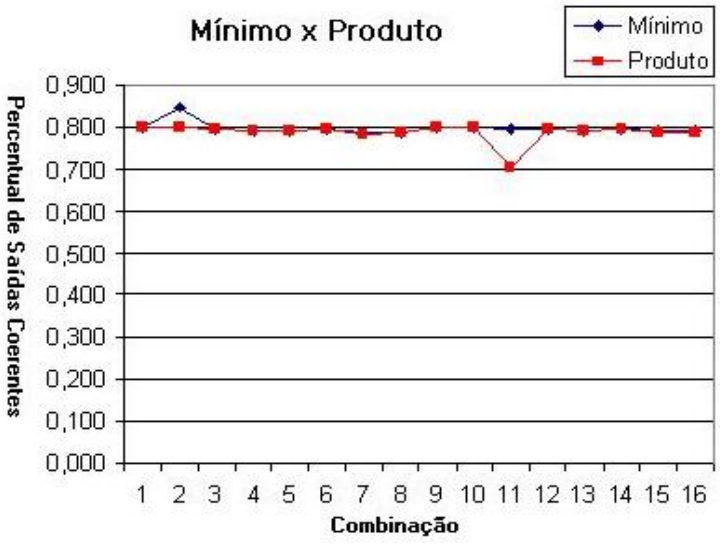


Figura 4.4: Comparação entre os dois métodos de composição.

Os gráficos individuais dos modelos de composição apresentados na Tabela 4.13 se encontram no Apêndice C (Figura C.1 para *mínimo* e C.2 para *produto*).

4.3.2 - Resultados Obtidos na Etapa de Desfuzificação

Ressalta-se que fez-se a desfuzificação para validar o método MDCAR em relação ao método tradicional de medida de atração-repulsão

Os resultados obtidos nesta etapa foram comparados com a medida de referência *lift* para verificar se as saídas obtidas estão próximas ou não da saída desejada. A diferença média entre a saída do MDCAR para cada um dos 128 testes e a saída de referência *lift* é calculada conforme (4.6). Por exemplo, suponha as entradas e saídas da Tabela 4.14, a diferença média é igual a 0,043.

$$DM = \frac{\sum_{i=1}^{n} |MDCAR_i - LIFT_i|}{n}, \qquad (4.6)$$

onde: n = número de linhas da tabela de resultados;

MDCAR – Método Difuso para Cálculo de Atração/Repulsão;

LIFT – Medida de Atração/Repulsão (GROTH, 2000).

Tabela 4.14: Exemplos da etapa de desfuzificação

Entradas			Saídas	
FRA	**FREAB**	**FROAB**	**LIFT**	**MDCAR**
0,87	0,70	0,73	0,03	0,06
0,57	0,10	0,11	0,01	0,02
0,78	0,31	0,16	-0,20	-0,20
0,50	0,455	0,450	-0,01	-0,02
0,67	0,55	0,18	-0,55	-0,42
0,74	0,40	0,53	0,17	0,09

Os termos apresentados na Tabela 4.14 têm os seguintes significados:

- FRA – Freqüência Relativa do antecedente da regra “se A então B”;
- FREAB – Freqüência Relativa Esperada de A e B;
- FROAB – Freqüência Relativa Obtida de A e B;
- LIFT – Medida de Atração/Repulsão (GROTH, 2000);
- MDCAR – Método Difuso para Cálculo de Atração/Repulsão.

As diferenças médias para todos os testes realizados são mostradas na Tabela 4.15, cujos campos têm os seguintes significados:

- MM – Modelo de *Mamdani* com composição *Min*;
- MP – Modelo de *Mamdani* com composição *Prod*;
- TKM – Modelo de *Takagi-Sugeno* com composição *Min*;
- TKP – Modelo de *Takagi-Sugeno* com composição *Prod*;
- TSM – Modelo de *Tsukamoto* com composição *Min*;
- TSP – Modelo de *Tsukamoto* com composição *Prod*.

Tabela 4.15: Diferença Média entre *Lift* e MDCAR

Combinação	MM	MP	TKM	TKP	TSM	TSP
1	0,124	0,123	0,094	0,094	0,100	0,115
2	0,124	0,123	0,094	0,096	0,101	0,115
3	0,134	0,127	0,096	0,097	0,094	0,105
4	0,134	0,128	0,098	0,098	0,096	0,107
5	0,127	0,125	0,087	0,088	0,122	0,138
6	0,126	0,124	0,089	0,091	0,127	0,141
7	0,136	0,131	0,089	0,090	0,115	0,127
8	0,136	0,131	0,092	0,093	0,120	0,130
9	0,126	0,124	0,095	0,095	0,101	0,115
10	0,126	0,124	0,098	0,098	0,103	0,117
11	0,136	0,130	0,097	0,098	0,094	0,107
12	0,136	0,130	0,100	0,100	0,099	0,110
13	0,139	0,127	0,088	0,089	0,125	0,139
14	0,139	0,126	0,090	0,092	0,131	0,143
15	0,129	0,134	0,090	0,092	0,119	0,130
16	0,129	0,135	0,093	0,094	0,222	0,134
Média	**0,131**	**0,128**	**0,093**	**0,094**	**0,117**	**0,123**
Desvio Padrão	**0,005**	**0,003**	**0,004**	**0,003**	**0,012**	**0,012**

Como pode ser observado na Tabela 4.15 e na Figura 4.5, o modelo de *Takagi-Sugeno* teve o melhor desempenho, apesar de nas combinações 3, 4 e 11 o modelo de *Tsukamoto* ter apresentado um melhor resultado. Em média, o modelo que apresenta o melhor resultado é o de *Takagi-Sugeno*, como é mostrado a seguir.

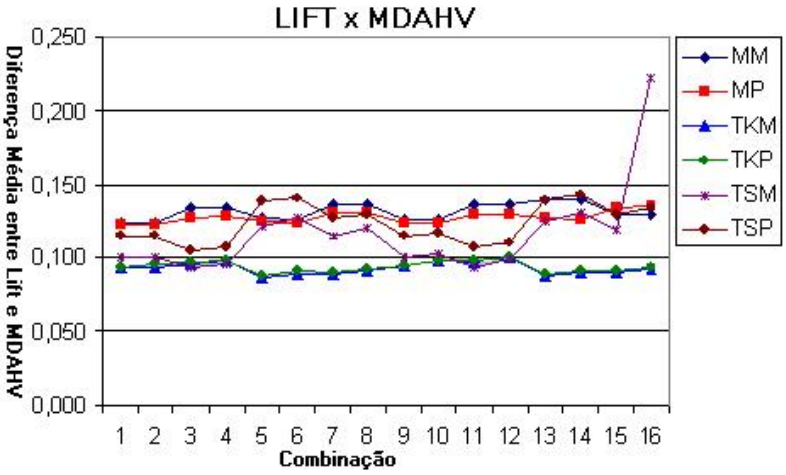


Figura 4.5: Diferenças Médias para os modelos testados.

A diferença média entre o modelo de Takagi-Sugeno (TKM – Takagi-Sugeno, *mínimo* e TKP – Takagi-Sugeno, *produto*) e os demais testados, é grande, como ilustra a Figura 4.6. Já a diferença entre *mínimo* e *produto* é pequena.

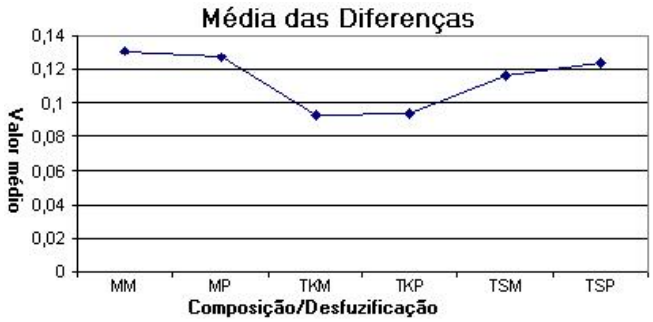


Figura 4.6: Média das diferenças entre os modelos testados.

As tabelas que deram origem à Tabela 4.15 e aos gráficos individuais se encontram no Apêndice C.

Pelos dados obtidos nesta pesquisa, constata-se que o melhor modelo de desfuzificação é o de *Takagi-Sugeno,* que pode ser usado com interseção: padrão e produto, porém optou-se por usar o segundo tipo de composição, para o método MDCAR proposto, visto que, em conjunto com o método de desfuzificação escolhido, apresenta resultado melhor.

Após a determinação do modelo ideal para composição do MDCAR, outros testes foram realizados a fim de fazer a sua validação. Os resultados desses testes são apresentados na próxima seção.

4.3.3 – Resultados Finais

Para fazer os últimos testes com o modelo escolhido para o MDCAR, foram selecionadas, por sorteio, 9.000 possíveis associações para cada base de dados: rede de supermercados, loja de departamentos e rede de livrarias. Os resultados são apresentados na Tabela 4.16, onde: as bases são numeradas de 1 a 3, para os três

segmentos de mercado; a coluna classificação é o resultado percentual para o número de saídas coerentes da etapa de classificação, conforme explicado anteriormente; a coluna desfuzificação é a diferença média entre a saída do MDCAR e a medida LIFT de referência.

Tabela 4.16: Resultado dos últimos testes

Base	Nº de casos	% de classificação adequada	Diferença média
1	3000	0,79310	0,09348
2	3000	0,79700	0,09207
3	3000	0,80080	0,09072

Como pode ser observado na Tabela 4.16, não há grande diferença entre os resultados obtidos com dados escolhidos aleatoriamente de diferentes segmentos do mercado e também entre os resultados obtidos anteriormente. Portanto, o método desenvolvido pode ser aplicado a qualquer um desses segmentos, sem necessidade de adequação.

Capítulo 5 – Considerações Finais

Neste capítulo são apresentadas, baseadas na literatura, nos testes e no desenvolvimento de um método para cálculo de atração/repulsão, as principais conclusões da pesquisa e são apontadas sugestões para pesquisas futuras.

5.1 - Conclusões

Os métodos tradicionais de busca por regras de associação usados no cálculo de uma medida de atração/repulsão não consideram a imprecisão contida na matriz de co-ocorrência, porque esta matriz que é composta por zeros e uns (presença ou não presença de determinados itens) não considera a intensidade da associação. Além disso, as medidas mais utilizadas não deixam claro se há atração ou repulsão, cabendo ao usuário a tarefa de interpretá-los.

Dessa forma, nesta pesquisa, foi proposto um método baseado na lógica difusa que pode tratar dessa imprecisão.

O MDCAR é um método que mapeia entradas numéricas para termos lingüísticos. Esse mapeamento é feito com objetivo de tratar a imprecisão contida na matriz de co-ocorrência que é usada para o calculo de atração/repulsão. As entradas são valores numéricos de freqüências e a saída é uma classificação de associação. Esta classificação pode ser de atração ou repulsão com grau de associação baixa, moderada ou alta.

Para isso, na modelagem do método proposto, foi necessário:

- o resgate junto à literatura especializada, de subsídios referentes a *Market Basket Analysis* (MBA) e à lógica difusa;
- identificação dos métodos mais usados em MBA e dos modelos de regras usados na lógica difusa;
- a construção de conjuntos difusos para representar termos lingüísticos, usados para as variáveis de entrada;

- a adequação dos limites dos intervalos das funções de pertinência, através de heurísticas.

Foram testadas várias combinações de funções de pertinência, com várias amostras de associações entre itens, oriundas de base de dados de três segmentos comerciais. Algumas dessas combinações apresentaram desempenhos razoáveis, mas a escolha pela combinação, cujas funções são não-lineares em vez de lineares, deveu-se ao fato de essa combinação apresentar alterações mais suaves na imagem, quando se caminha de um ponto a outro do domínio.

O modelo difuso de *Takagi-Sugeno* foi escolhido por apresentar os melhores resultados para todas as combinações de funções testadas, ou seja, foi o modelo que se mostrou mais robusto durante os testes. Os resultados obtidos se apresentaram mais adequados aos apresentados por outros modelos. Além disso, é o modelo que fez com que o método proposto, ao ser comparado com o método tradicional de cálculo de atração/repulsão, fornecesse saídas próximas das esperadas pelo método tradicional.

O MDCAR mostrou bons resultados e pode ser aplicado à área comercial para análise de dados históricos de vendas. Além disso, pode ser usado nos pontos de vendas para auxiliar o atendente a oferecer um novo produto a determinados clientes, baseado na sua compra atual; porque a resposta do sistema pode ser dada em linguagem natural, o que torna acessível a qualquer usuário do sistema.

Pode-se usar o método também para fazer consultas usando linguagem natural. Isto depende da criatividade de cada analista ao implementar o sistema que usará o método. A vantagem de usar o MDCAR em sistemas comerciais é grande, pois não há necessidade de fazer grandes alterações no sistema comercial atual da empresa. Basta criar um módulo que o implemente e fazer que esse módulo e o sistema se comuniquem. Isso representa uma grande economia para a empresa.

5.2 – Trabalhos Futuros

Considerando que o método pode ser aperfeiçoado, sugere-se:

- refinar a base de regras, acrescentando, modificando ou retirando regras para detectar possíveis falhas na base de regras;
- redefinir os conjuntos para cada variável de entrada, acrescentando mais variáveis lingüísticas e verificar se isso torna a base de regras mais consistente;
- testar o modelo em outras áreas do conhecimento, como por exemplo na medicina para verificar a associação entre doenças;
- trabalhar com outras variáveis de entrada, tais como suporte e confiança para mapear em uma única saída, ou seja, aproximar uma função para essas entradas.

Além destas sugere-se também:

- desenvolver novas ferramentas “inteligentes” para agilizar o processo de pré-processamento dos dados, ou seja, na preparação dos dados para o processo de descoberta de conhecimento;
- unificar algumas técnicas de mineração de dados, visto que às vezes são necessárias várias destas para conseguir atingir o objetivo.

Referências

AGGARWAL, C. and YU, P. **Online generation of association rules**. *ICDE-98*, 1998, pp. 402-411.

AGRAWAL, R., IMIELINSKI, T., SWAMI, A. **Mining association rules between sets of items in large databases**. *SIGMOD-1993*, 1993, pp. 207-216.

AGRAWAL, R. and SRIKANT, R. **Fast algorithms for mining association rules**. *VLDB-94*, 1994.

AMARAL, Fernanda C. N. **Mineração de dados**: Técnicas e Aplicações para o Marketing Direto. São Paulo: Berkeley, 2001.

ARIA, Massino; MOLA, Francisco; SICILIANO, Roberta. **Growing and Visualizing Prediction Paths Trees in Market Basket Analysis**. In: XV Conference on Computational Statistics, Berlin - Germany, 2002.

BARBETTA, Pedro A. **Estatística Aplicada às Ciências Sociais**. Florianópolis: Editora da UFSC, 2004.

BARRETO, Jorge M. **Inteligência Artificial no Limiar do Século XXI**. Florianópolis: J. M. Barreto, 2000.

BERK, Kenneth N. **Data Analysis with Systat**. USA: SYSTAT, 1994.

BERRY, Michael, J. A. & LINOFF, Gordon. **Mineração de dados Techniques**: for Marketing, Sales, and Customers. USA: John Wiley & Sons, 1997.

BRIN, S. MOTWANI, R. ULLMAN, J. and TSUR, S. **Dynamic Itemset counting and implication rules for market basket data**. *SIGMOD-97*, 1997, pp. 255-264.

CARVALHO, Luís A. V. de. **Datamining**: A Mineração de Dados no Marketing, Medicina, Economia, Engenharia e Administração. São Paulo: Érica, 2001.

CURY, Marcus V. Q. Método para Classificar o Desempenho de Sistemas de Transporte Urbano com Uso da Lógica difusa. **Revista Transporte – ANPET**, Brasília, Abril de 2003.

DEVLIN, Keith. **Logic and Information**. England: Cambridge University Press, 1991.

DRAESEKE, Robert; GILES, David.E.A., **A Fuzzy Logic Approach to Modelling the Underground Economy**. In: International Conference on Modelling and Simulation

(MODISM 99) - Modelling and Simulation Society of Australia and New Zealand. Vol. 2, p. 453-458, December, 1999.

DUALIBE, Carlos; JESPERS, Paul; VERLEYSEN, Michel. **On Designing Mixed-Signal Programmable Fuzzy Logic Controllers as Embedded Subsystems in Standard CMOS Technologies.** In: 14th Symposium on Integrated Circuits and Systems. Brasília - DF, p. 194-200, 10 a 15, September, 2001.

FAYYAD, Usama, PIATETSKY-SHAPIRO, Gregory e SMYTH Padhraic. **The DCBD Process for Extracting Useful Knowledge.** Communications of the ACM Digital Library, Novembro, 1996, Vol. 39, p 27-34.

GABBAY, Dov M. **What is Logical System?** USA: Oxford Science Publications, 1994.

GROTH, Robert. **Data Mining**: Building Competitive Advantage. New Jersey – USA: Prentice Hall, 2000.

GUIMARÃES, Márcio G. **Um Sistema de Apoio à Dosimetria da Pena do Código Penal Brasileiro Utilizando Fuzzy Logic**. Florianópolis, 2003, 106p. Dissertação (Mestrado em Ciência da Computação) – Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Santa Catarina.

HAN, J. and FU, Y. **Discovery of multiple-level association rules from large databases**. *VLDB-95*.

HAN, Jiawei & KAMBER, Micheline. **Mineração de dados**: Concepts and Techniques. USA: Morgan Kaufmann, 2001.

KANDEL, Abraham. **Fuzzy Mathematical Techiniques with Applications**. USA: Addison-Wesley Publishing, 1986.

KLIR, George; YUAN, Bo. **Fuzzy Sets and Fuzzy Logic**: Theory and Applications. USA: Prentice Hall, 1995.

KLÖSGEN, Willi & ZYTKOW, Jan M. **Handbook of Mineração de dados and Knowledge Discovery**. USA: Oxford University Press, 2002.

KOSKO, Bart. **Fuzzy Engineering**. USA: Prentice-Hall, 1997.

LAKSHMANAN, L.; NG. R. T.; HAN, J. **Exploratory mining and pruning optimizations of constrained association rules**. *SIGMOD-98*, 1998.

MATTHEWS, Chris. **Fuzzy Concepts and Formal Methods: A Sample Specification for a Fuzzy Expert System**. In: World Congress on Computational Intelligence (WCCI 2002), IEEE Press, 2002.

MENDES, Ilza M. B. **Regras de Associação Negativas**. Niterói-RJ, 2002, 63p. Dissertação (Mestrado em Computação Aplicada e Automação) – Programa de Pós-Graduação em Computação Aplicada e Automação da Universidade Federal Fluminense.

NISSANKE, Nimal. **Introductory Logic and Sets for Computer Scientists.** England: Addison Wesley Longman, 1999.

NOLT, J.; ROHATYN, D. **Lógica**. São Paulo: Mcgraw-Hill, Inc., 1991.

NOTARI, Daniel L. **Aplicação de Redes Neurais Artificiais à Mineração de Dados**. Disponível em:<http://www.inf.ufrgs.br/~dlnotari/trabalhos/ucs/arnmd/index.html>. Acesso em 15/02/2000.

ORTEGA, Neli R. S. **Aplicações da Teoria de Conjuntos difusos a Problemas da Biomedicina**. São Paulo-SP, 2001, 166p. Tese (Doutorado em Ciências) – Programa de Pós-Graduação Ciências da Universidade de São Paulo.

PARK, J. S. CHEN, M. S. and YU, P. S. **An effective hash based algorithm for mining association rules**. *SIGMOD-95*, 1995, pp. 175-186.

RASTOGI, R. and SHIM, K. **Mining optimized association rules with categorical and numeric attributes**. *ICDE –98*.

RESSOM, H.; REYNOLDS, R.; VARGHESE, R. S. Increasing the efficiency of fuzzy logic-based gene expression data analysis**. Physiological Genomics**, Stanford University - CA, v. 2, n. 13, p. 107-117, 16 de Abril de 2003.

RIBEIRO, Rita A. & MOREIRA, Ana M. Fuzzy Query Interface for a Business Database. **International Journal of Human-Computer Studies,** V. 58 , N. 4 p. 363-391, Abril de 2003.

ROBIN, Jacques; BEZERRA, Ricardo. **Descoberta de Conhecimento em BD**. Disponível em: <www.di.ufpe.br/~compint/aulas-IAS/DCBD-991/DCBD.ppt >. Acesso em: 13/08/2003.

ROSS, Timothy J. **Fuzzy Logic with Engeneering Applications.** USA: McGraw-Hill, 1995.

ROYES, Gleiber F. **Plataforma Híbrida Fuzzy-Multicritério-RBC para o Apoio à Análise de Políticas**. Florianópolis, 2003, 195p. Tese (Doutorado em Ciência da Computação) – Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Santa Catarina.

SANDRI, Sandra; CORREA, Cláudio. **Lógica Nebulosa**. In: V Escola de Redes Neurais, Promoção - Conselho Nacional de Redes Neurais. p. c073-c090, 19 de julho- ITA, São José dos Campos - SP, 1999.

SAVASARE, A.; OMIECINSKI, E.; NAVATHE, S. **Mining for Strong Negative Associations in a Large Database of Customer Transactions**. In: 14th International Conference on Data Engineering, Florida, 494-502, 1998.

SCREMIN, Marcos A. A. **Método para a Seleção do Número de Componentes Principais com Base na Lógica Difusa**. Florianópolis, 2003, 124p. Tese (Doutorado em Engenharia de Produção) – Programa de Pós-Graduação em Engenharia de Produção Universidade Federal de Santa Catarina.

SOUZA, Flávio J. de. **Modelos Neuro-Fuzzy Hierárquicos**. Tese de Doutorado, DEE-PUC/RJ, 29 de abril de 1999.

SRIKANT, R., VU, Q. and AGRAWAL, R. **Mining association rules with item constraints**. *KDD-97*, 1997, pp. 67-73.

STURM, Ulrike et al. **Análise da Ocupação em Áreas de Preservação Permanente na Área Urbana do Município de Matinhos Utilizando a Imagem Ikonos II**. In: III Colóquio Brasileiro de Ciências Geodésicas – Novos Desenvolvimentos em Ciências Geodésicas, Curitiba, 06 a 09 de maio de 2003.

TANSCHEIT, Ricardo. **Lógica difusa, Raciocínio Aproximado e Mecanismo de Inferência**. Disponível em: <http://www.ica.ele.puc-rio.br/cursos/download/LN-Logica_Controle_Fuzzy.pdf>. Acesso em: 28/08/2003.

YAGER, R. R. et al. **Fuzzy Sets and Applications**: Selected Papers by L. A. Zadeh. USA: John Willey & Sons, 1987.

ZADEH, L. A., **Fuzzy Sets**, Inf. Control 8, 338-353, 1965.

VELOSO, A. et. al. **Mineração Incremental de Regras de Associação**. In: XVI Simpósio Brasileiro de Banco de Dados. IME – Rio de Janeiro, p. 80-94, 2001.

WOOLF, Peter J.; WANG, Yixin. A fuzzy logic approach to analyzing gene expression data. **Physiological Genomics**, Stanford University - CA, v. 1, n. 3, p. 9-15, Abril, 2000.

Apêndice A: Conteúdo do CD em Anexo.

Neste apêndice é descrito o conteúdo do CD que segue em anexo a esta tese.

Os arquivos do CD são os seguintes, conforme ilustrado pela Figura A.1:

- BaseTrab.dbf – arquivo com 1.242.507 linhas, que é a base que foi usada para desenvolvimento do método;
- BaseTest.dbf – arquivo com 124.200 linhas, que é a base usada para os testes;
- pastas, testes e testes_finais, que contém os resultados dos testes.

Figura A.1: Conteúdo do CD.

Como pode ser observado pela Figura A.1, a pasta testes contém as subpastas: classificação, *Mamdani*, *Takagi* e *Tsukamoto*. Cada subpasta dessas contém as subpastas *Min* e *Prod*, são nessas pastas que estão armazenados os arquivos com os resultados dos testes, conforme o exemplo mostrado na Tabela A.1. A pasta *Min* contém os resultados para a composição *Min* e a pasta *Prod* para a composição *Prod*. Já, na subpasta testes_finais, tem-se takagi_min, resultados dos testes finais com composição *Min* e takagi_prod, testes finais com composição *Prod*.

Tabela A.1: Exemplo dos resultados obtidos nos testes

FRA	FREAB	FROAB	LIFT	NUMÉRICO	MDCAR
0,65422	0,28036	0,17176	-0,16600	-0,19427	RMODERADA
0,64334	0,50618	0,27941	-0,35249	-0,28280	RMODERADA
0,28371	0,15562	0,24790	0,32526	0,26784	AALTA
0,79312	0,25060	0,11049	-0,17666	-0,17244	RMODERADA
0,81741	0,50052	0,06210	-0,53635	-0,37856	RALTA
0,54524	0,07441	0,06150	-0,02368	-0,02717	RBAIXA
0,55805	0,31772	0,19111	-0,22688	-0,24645	RMODERADA
0,55765	0,18716	0,13970	-0,08511	-0,05827	RBAIXA
0,72341	0,18737	0,13773	-0,06862	-0,05355	RBAIXA
0,85751	0,40004	0,28230	-0,13730	-0,13100	RBAIXA
0,69052	0,16416	0,10704	-0,08272	-0,06083	RBAIXA
0,34493	0,09000	0,23811	0,42939	0,50989	AALTA
0,50273	0,44738	0,16734	-0,55704	-0,48031	RALTA
0,73542	0,43967	0,45985	0,02744	0,03636	ABAIXA
0,80812	0,39478	0,21604	-0,22118	-0,19369	RBAIXA

Há 128 tabelas (16 combinações de funções multiplicado por 4 modelos multiplicado por 2 métodos de composição) com 124.200 linhas cada, como a do exemplo da Tabela A.1, exceto as tabelas resultantes do modelo de classificação que não contém a coluna *NUMÉRICO*. Cada tabela dessas está em um arquivo com nomes como “ZEPISIG.dbf”, que significa que a combinação de função usada foi Z, *Pi* e Sigmoidal.

As colunas, Tabela A.1 FRA, FRE e FRO significam Freqüência Relativa de A, Freqüência Relativa Esperada de A e B e Freqüência Relativa Obtida de A e B; a coluna LIFT, valor da saída de referência; a coluna NUMÉRICO, o valor de saída *Numérico* do método e a coluna MDCAR é a saída em linguagem natural do método.

Apêndice B: **Funções de pertinência Utilizadas na Pesquisa**.

Neste apêndice são apresentadas as funções de pertinência usadas para a realização dos testes.

As figuras de B.1 a B.8 mostram os gráficos de cada função utilizada para fazer os experimentos.

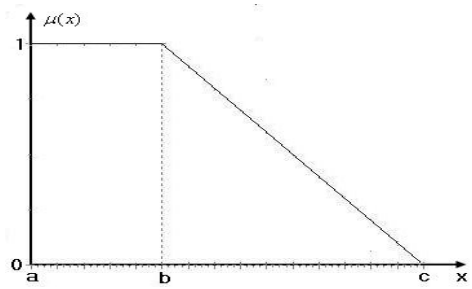


Figura B.1: Função L (TD).

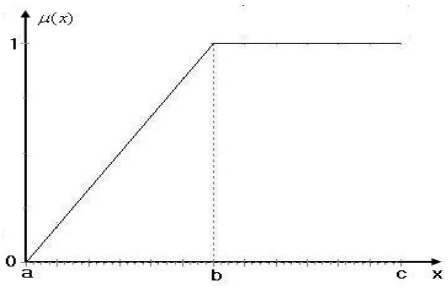


Figura B.2: Função Gama (TE).

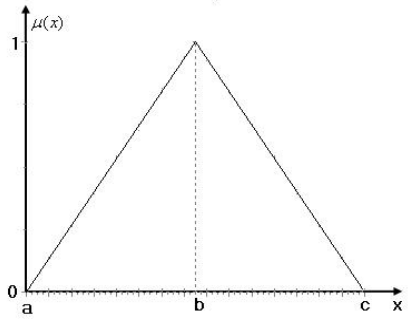


Figura B.3: Função triangular.

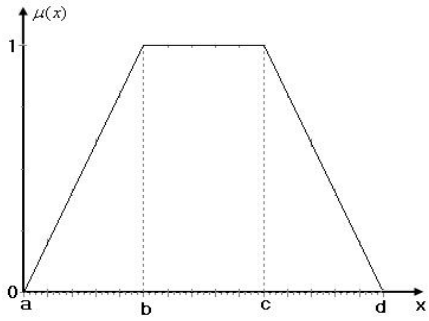


Figura B.4: Função trapezoidal.

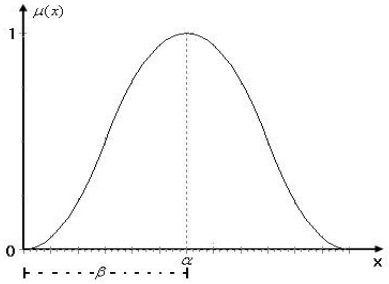


Figura B.5: Função PI.

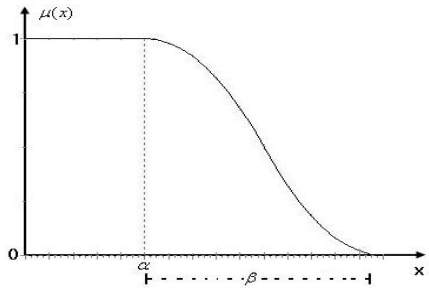


Figura B.6: Função Z.

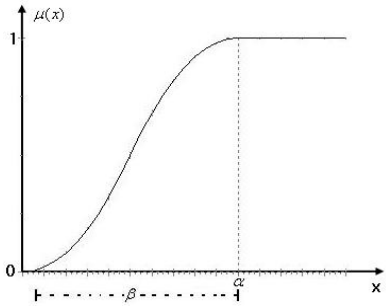


Figura B.7: Função Sigmoidal.

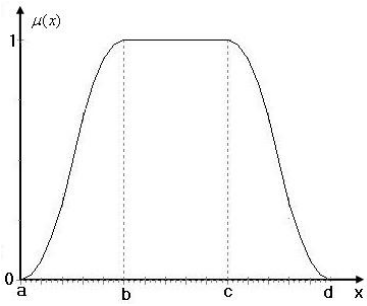


Figura B.8: Função sino.

A Tabela B.1 mostra os resultados obtidos para cada conjunto de parâmetros, com as entradas FRA = 0,52, FRE = 0,36 e FRO = 0,28.

Tabela B.1: Intervalos usados para os testes iniciais, valores entre 0 e 100

Conjuntos							
Baixa		**Moderada**		**Alta**		**Resultado**	
Função	**Intervalo**	**Função**	**Intervalo**	**Função**	**Intervalo**	**MDCAR**	***Numérico***
L(TD)	0,10,40	Triangular	10,40,60	TE	40,70,100	**Rmod**	-0,06
	0,15,45		15,45,65		45,75,100	**Abaixa**	-0,37
	0,10,35		10,35,55		35,65,100	**Amod**	-0,09
	0,10,30		10,35,60		35,70,100	**Abaixa**	-0,07
	0,10,40		**10,40,70**		**40,70,100**	**Rbaixa**	**-0,11**
	0,10,30		10,30,60	Sig	70,36	**Rmod**	-0,07
	0,10,30		10,30,60		70,40	**Abaixa**	-0,05
	0,10,30		10,30,60		70,45	**Amod**	0,04
	0,10,30		10,30,60		60,41	**Abaixa**	-0,10
	0,10,40		**10,40,70**		**70,30**	**Rbaixa**	**-0,12**
	0,10,30	Trapézio	10,30,40,60	TE	30,60,100	**Amod**	-0,11
	0,10,30		10,20,30,60		30,60,100	**Abaixa**	-0,11
	0,10,40		**10,35,45,70**		**40,70,100**	**Rbaixa**	**-0,11**
	0,10,40		**10,35,45,70**	**Sig**	**70,30**	**Rbaixa**	**-0,13**
	0,10,30	Pi	35,15	TE	30,60,100	**Rmod**	-0,09
	0,10,40		**40,30**		**40,70,100**	**Rbaixa**	**-0,11**
	0,10,30		30,15	Sig	60,30	**Abaixa**	-0,10
	0,10,30		30,18		60,30	**Amod**	-0,12
	0,10,30		30,19		60,30	**Abaixa**	-0,12
	0,10,40		**40,30**		**70,30**	**Rbaixa**	**-0,12**
	0,10,30	Sino	10,18,38,60	TE	30,60,100	**Rmod**	-0,11
	0,10,30		10,19,39,60		30,60,100	**Abaixa**	-0,11
	0,10,30		10,29,39,60		30,60,100	**Amod**	-0,11
	0,10,30		10,30,40,60		30,60,100	**Abaixa**	-0,11
	0,10,30		10,35,55,75		30,60,100	**Rbaixa**	0,13
	0,10,30		10,30,50,70		30,60,100	**Rbaixa**	0,15
	0,10,30		10,30,50,60		30,60,100	**Rbaixa**	0,15
	0,10,30		10,25,45,70		30,60,100	**Amod**	-0,11
	0,10,30		10,25,45,60		30,60,100	**Abaixa**	-0,11
	0,10,40		**10,35,45,70**		**40,70,100**	**Rbaixa**	**-0,11**
	0,10,40		**10,35,45,70**	Sig	**70,30**	**Rbaixa**	**-0,13**
Z	0,15,35	Triangular	10,30,60	TE	30,60,100	**Abaixa**	-0,11
	0,10,40		**10,40,70**		**40,70,100**	**Rbaixa**	**-0,11**
	0,10,40		**10,40,70**	Sig	**70,30**	**Rbaixa**	**-0,12**
	0,10,40	Trapézio	**10,35,45,70**	TE	**40,70,100**	**Rbaixa**	**-0,11**
	0,10,40		**10,35,45,70**	Sig	**70,30**	**Rbaixa**	**-0,13**
	0,10,40	Pi	**40,30**	TE	**40,70,100**	**Rbaixa**	**-0,11**
	0,10,40		**40,30**	Sig	**70,30**	**Rbaixa**	**-0,12**
	0,10,40	Sino	**10,35,45,70**	TE	**40,70,100**	**Rbaixa**	**-0,11**
	0,10,40		**10,35,45,70**	Sig	**70,30**	**Rbaixa**	**-0,13**

Apêndice C: Gráficos e Tabelas Obtidas Durante os Experimentos.

Neste apêndice é apresentado um conjunto de gráficos e tabelas que foram empregados ao longo da documentação.

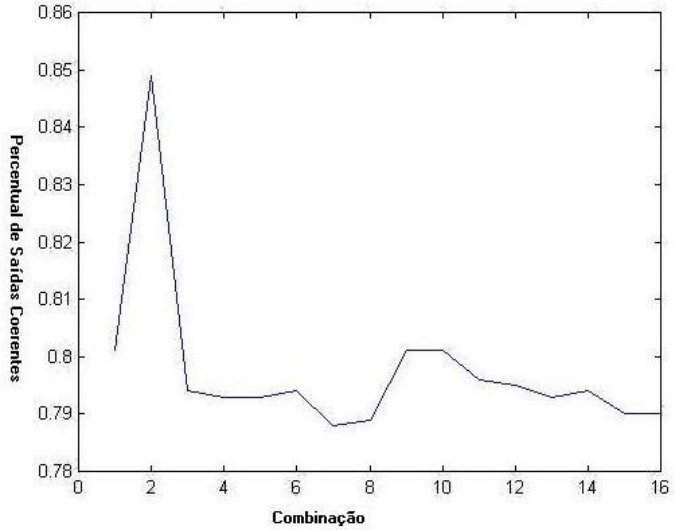


Figura C.1: Resultados da etapa de classificação para composição *Min*.

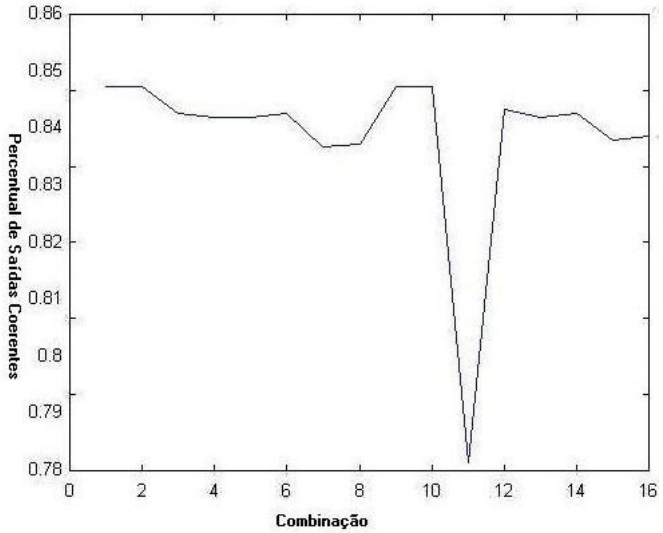


Figura C.2: Resultados da etapa de classificação para composição *Prod*.

Os campos das tabelas apresentadas a seguir têm os seguintes significados:

- TABELA – tabela usada para o teste que se encontra no CD em anexo e é descrito no Apêndice A;

- DIFNUMÉRICO – diferença entre a saída *numérico* do MDCAR e a medida de referência, *lift*.

Tabela C.1: Resultados do modelo *Mamdani* com composição *Min*

TABELA	DIFNUMÉRICO
TDTRITE.dbf	0,124
TDTRISIG.dbf	0,124
TDTRATE.dbf	0,134
TDTRASIG.dbf	0,134
TDPITE.dbf	0,127
TDPISIG.dbf	0,126
TDSITE.dbf	0,136
TDSISIG.dbf	0,136
ZETRITE.dbf	0,126
ZETRISIG.dbf	0,126
ZETRATE.dbf	0,136
ZETRASIG.dbf	0,136
ZEPITE.dbf	0,139
ZEPISIG.dbf	0,139
ZESITE.dbf	0,129
ZESISIG.dbf	0,129

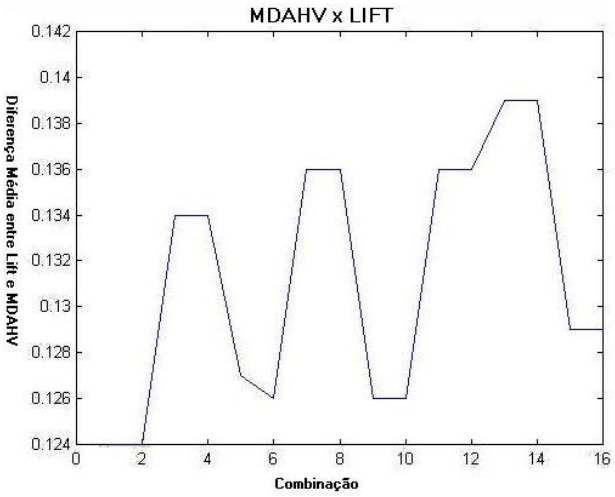


Figura C.3: Resultados do modelo *Mamdani* com composição *Min*.

Tabela C.2: Resultados do modelo *Mamdani* com composição *Prod*

TABELA	DIFNUMÉRICO
TDTRITE.dbf	0,123
TDTRISIG.dbf	0,123
TDTRATE.dbf	0,27
TDTRASIG.dbf	0,128
TDPITE.dbf	0,125
TDPISIG.dbf	0,124
TDSITE.dbf	0,131
TDSISIG.dbf	0,131
ZETRITE.dbf	0,124
ZETRISIG.dbf	0,124
ZETRATE.dbf	0,130
ZETRASIG.dbf	0,130
ZEPITE.dbf	0,127
ZEPISIG.dbf	0,126
ZESITE.dbf	0,134
ZESISIG.dbf	0,135

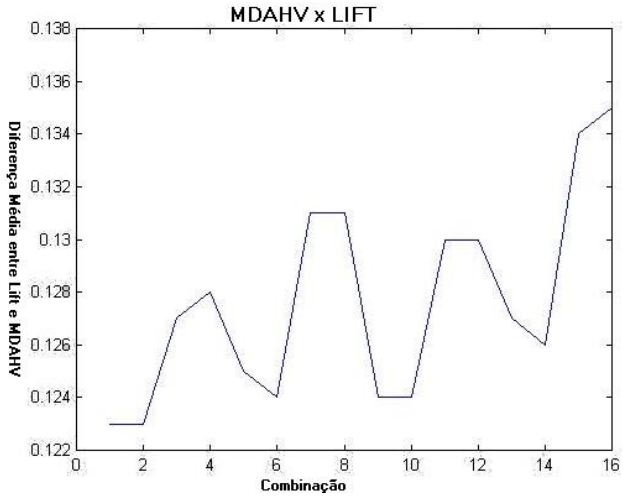


Figura C.4: Resultados do modelo *Mamdani* com composição *Prod*.

Tabela C.3: Resultados do modelo *Takagi-Sugeno* com composição *Min*

TABELA	DIFNUMÉRICO
TDTRITE.dbf	0,094
TDTRISIG.dbf	0,094
TDTRATE.dbf	0,096
TDTRASIG.dbf	0,098
TDPITE.dbf	0,087
TDPISIG.dbf	0,089
TDSITE.dbf	0,089
TDSISIG.dbf	0,092
ZETRITE.dbf	0,095
ZETRISIG.dbf	0,098
ZETRATE.dbf	0,097
ZETRASIG.dbf	0,100
ZEPITE.dbf	0,088
ZEPISIG.dbf	0,090
ZESITE.dbf	0,090
ZESISIG.dbf	0,093

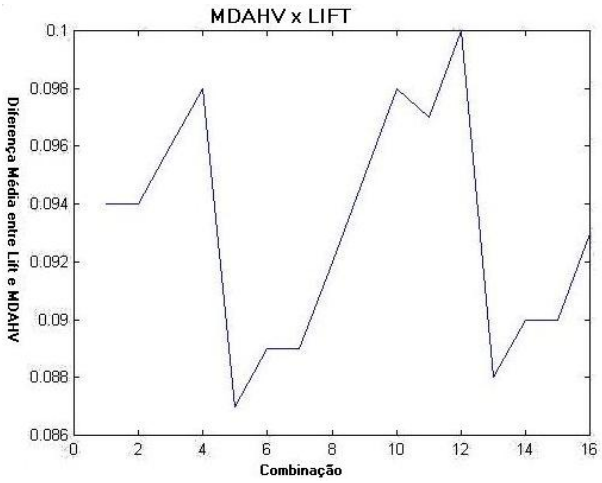


Figura C.5: Resultados do modelo *Takagi-Sugeno* com composição *Min*.

Tabela C.4: Resultados do modelo *Takagi-Sugeno* com composição *Prod*

TABELA	DIFNUMÉRICO
TDTRITE.dbf	0,094
TDTRISIG.dbf	0,096
TDTRATE.dbf	0,097
TDTRASIG.dbf	0,098
TDPITE.dbf	0,088
TDPISIG.dbf	0,091
TDSITE.dbf	0,090
TDSISIG.dbf	0,093
ZETRITE.dbf	0,095
ZETRISIG.dbf	0,098
ZETRATE.dbf	0,098
ZETRASIG.dbf	0,100
ZEPITE.dbf	0,089
ZEPISIG.dbf	0,092
ZESITE.dbf	0,092
ZESISIG.dbf	0,094

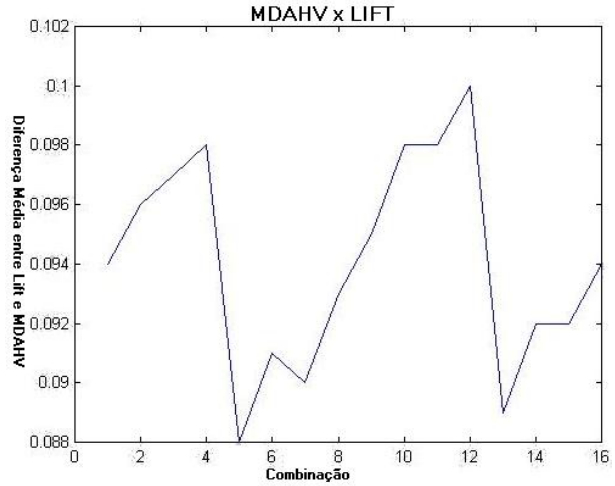


Figura C.6: Resultados do modelo *Takagi-Sugeno* com composição *Prod*.

Tabela C.5: Resultados do modelo *Tsukamoto* com composição *Min*

TABELA	DIFNUMÉRICO
TDTRITE.dbf	0,100
TDTRISIG.dbf	0,101
TDTRATE.dbf	0,094
TDTRASIG.dbf	0,096
TDPITE.dbf	0,122
TDPISIG.dbf	0,127
TDSITE.dbf	0,115
TDSISIG.dbf	0,120
ZETRITE.dbf	0,101
ZETRISIG.dbf	0,103
ZETRATE.dbf	0,094
ZETRASIG.dbf	0,099
ZEPITE.dbf	0,125
ZEPISIG.dbf	0,131
ZESITE.dbf	0,119
ZESISIG.dbf	0,222

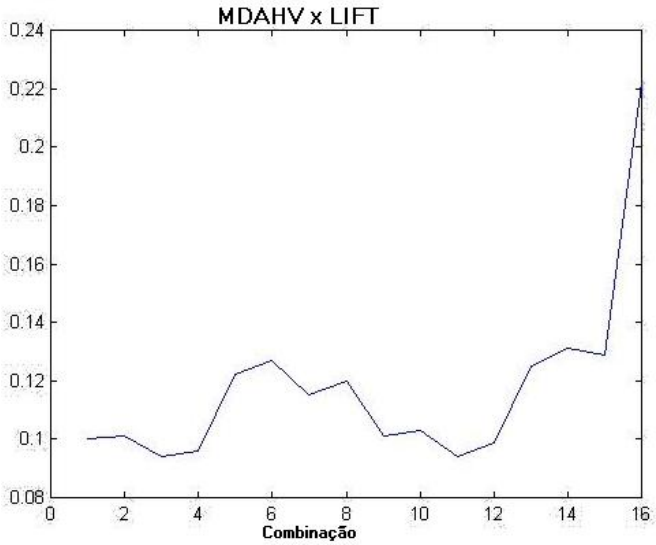


Figura C.7: Resultados do modelo *Tsukamoto* com composição *Min*.

Tabela C.6: Resultados do modelo *Tsukamoto* com composição *Prod*

TABELA	DIFNUMÉRICO
TDTRITE.dbf	0,115
TDTRISIG.dbf	0,115
TDTRATE.dbf	0,105
TDTRASIG.dbf	0,107
TDPITE.dbf	0,138
TDPISIG.dbf	0,141
TDSITE.dbf	0,127
TDSISIG.dbf	0,130
ZETRITE.dbf	0,115
ZETRISIG.dbf	0,117
ZETRATE.dbf	0,107
ZETRASIG.dbf	0,110
ZEPITE.dbf	0,139
ZEPISIG.dbf	0,143
ZESITE.dbf	0,130
ZESISIG.dbf	0,134

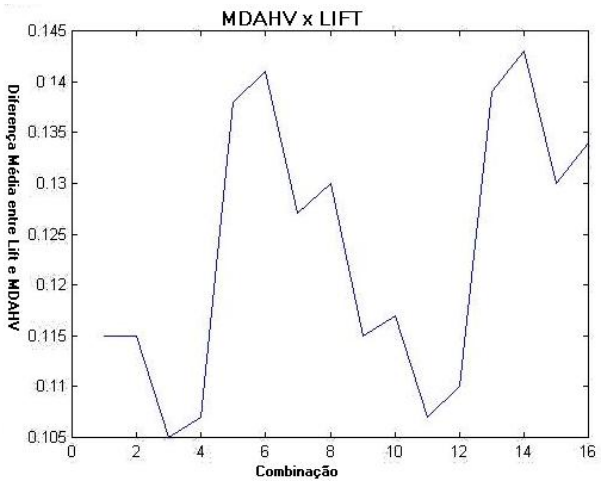


Figura C.8: Resultados do modelo *Tsukamoto* com composição *Prod*.

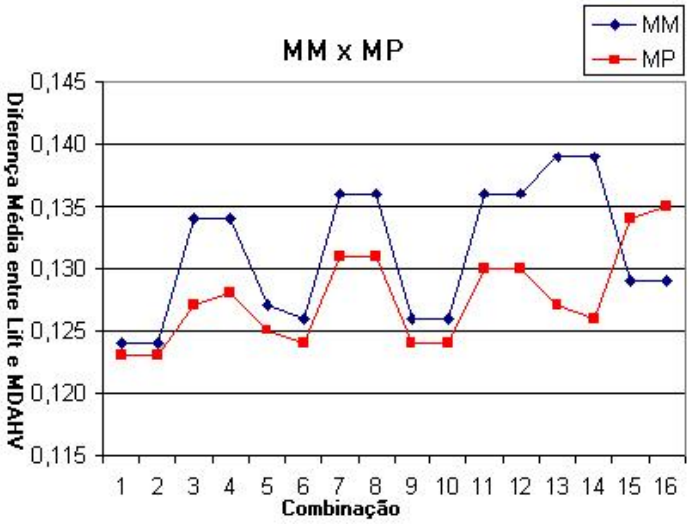


Figura C.9: Comparação entre composição *Min* e *Prod* para o modelo de *Mamdani*.

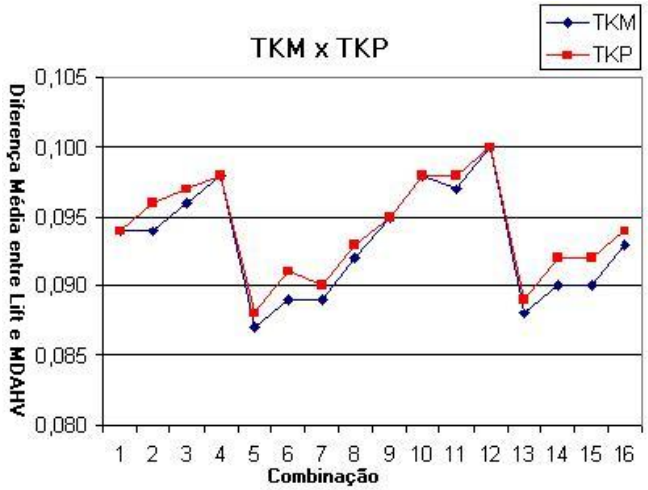


Figura C.10: Comparação entre composição *Min* e *Prod* para o modelo de *Takagi-Sugeno*.

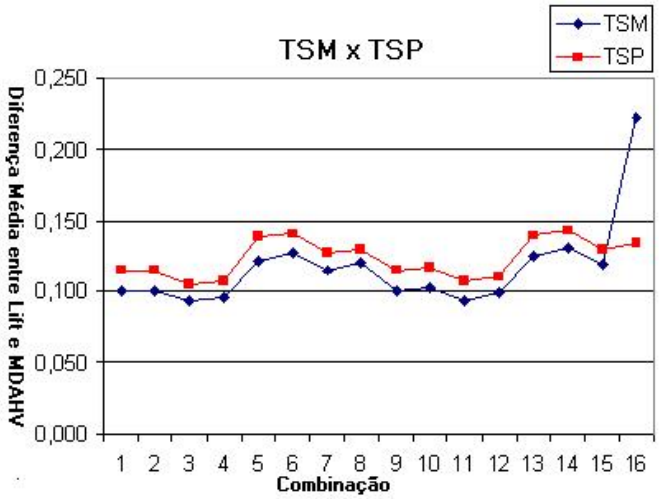


Figura C.11: Comparação entre composição *Min* e *Prod* para o modelo de *Tsukamoto*.

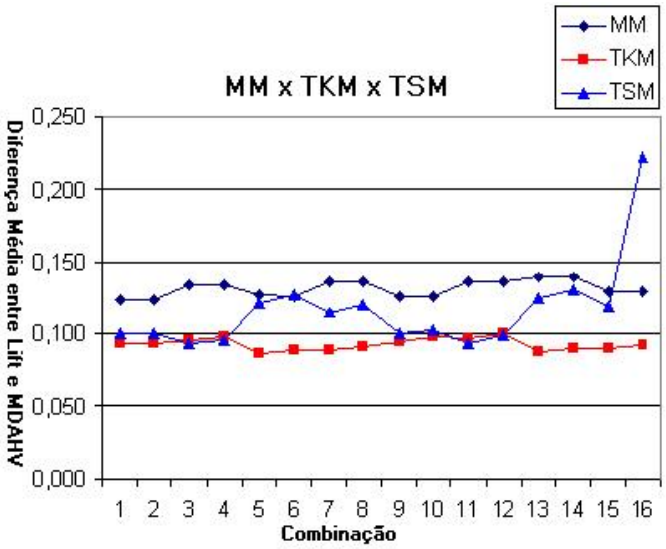


Figura C.12: Comparação entre *Mamdani, Takagi-Sugeno* e *Tsukamoto* com composição *Min*.

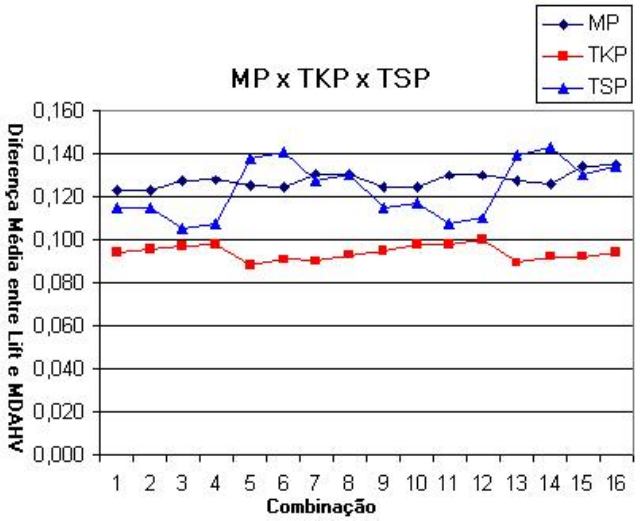


Figura C.13: Comparação entre *Mamdani, Takagi-Sugeno* e *Tsukamoto* com composição *Prod*.

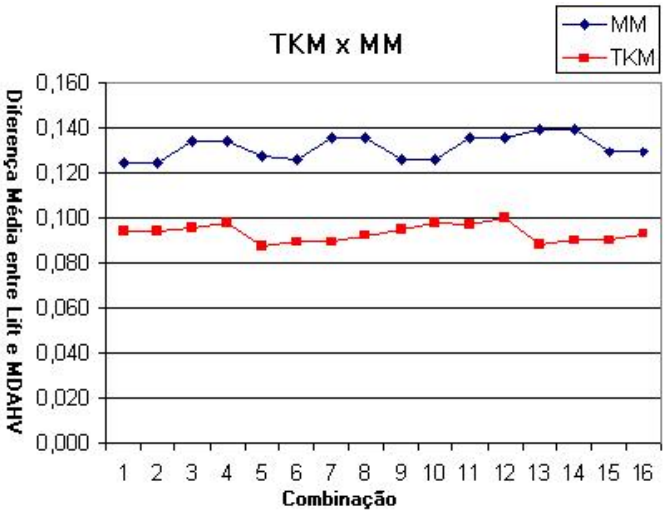


Figura C.14: Comparação entre *Mamdani* e *Takagi-Sugeno* com composição *Min*.

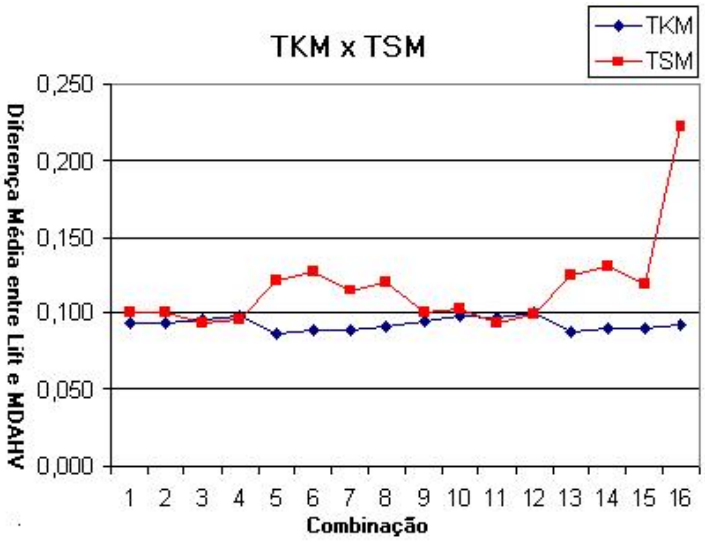


Figura C.15: Comparação entre *Tsukamoto* e *Takagi-Sugeno* com composição *Min*.

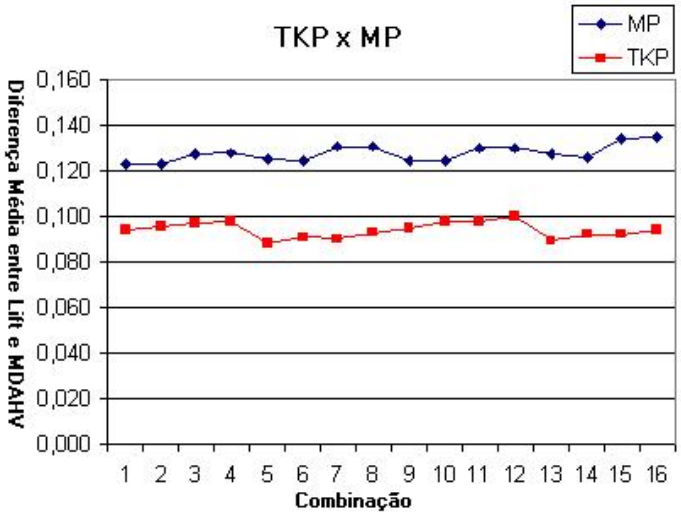


Figura C.16: Comparação entre *Mamdani* e *Takagi-Sugeno* com composição *Prod*.

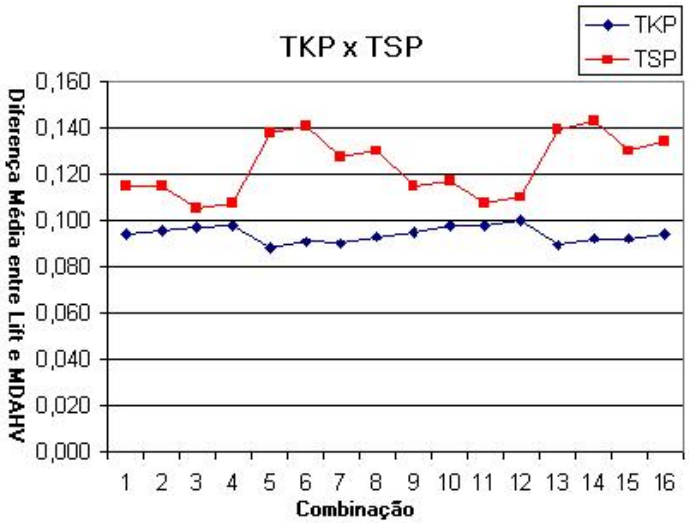


Figura C.17: Comparação entre *Tsukamoto* e *Takagi-Sugeno* com composição *Prod*.

Apêndice D: Gráficos das Combinações de Funções Usadas nesta Pesquisa

Neste apêndice são apresentados os gráficos das combinações de funções utilizadas durante os testes realizados nesta pesquisa.

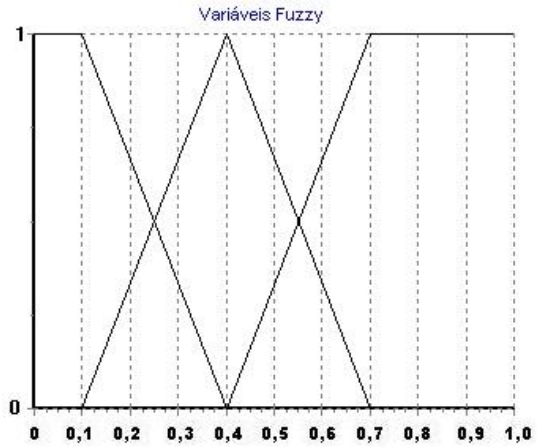


Figura D.1: Combinação 1 (Ver Tabela 4.6).

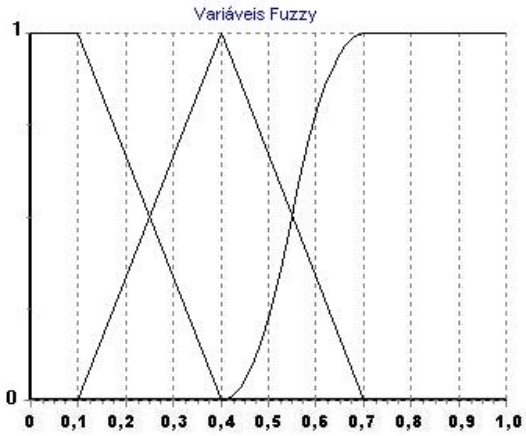


Figura D.2: Combinação 2 (Ver Tabela 4.6).

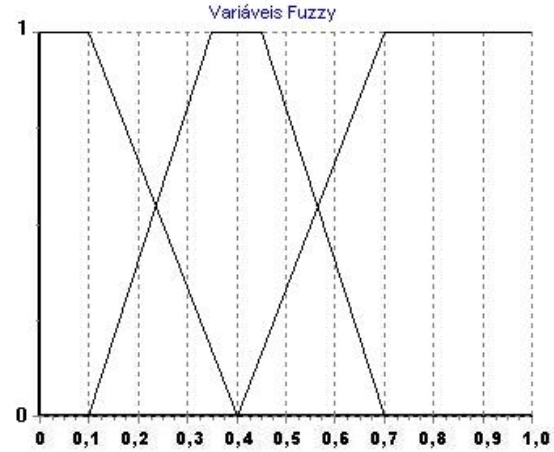


Figura D.3: Combinação 3 (Ver Tabela 4.6).

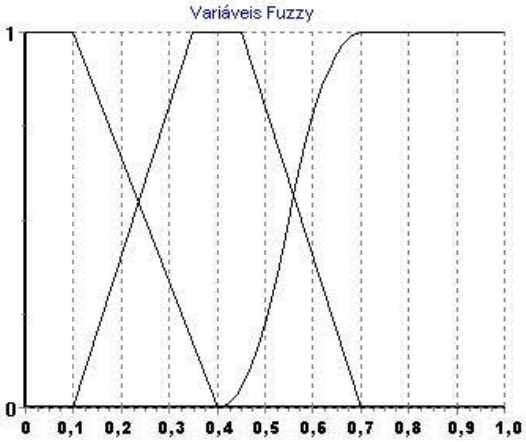


Figura D.4: Combinação 4 (Ver Tabela 4.6).

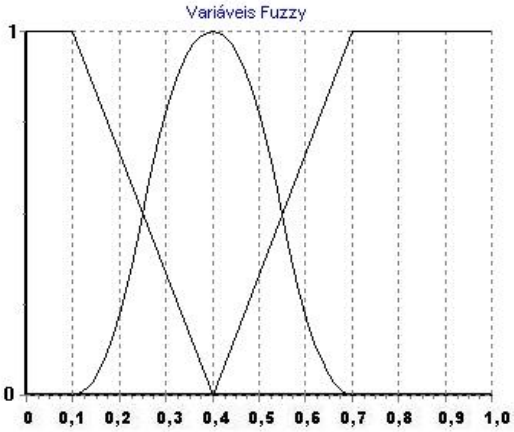


Figura D.5: Combinação 5 (Ver Tabela 4.6).

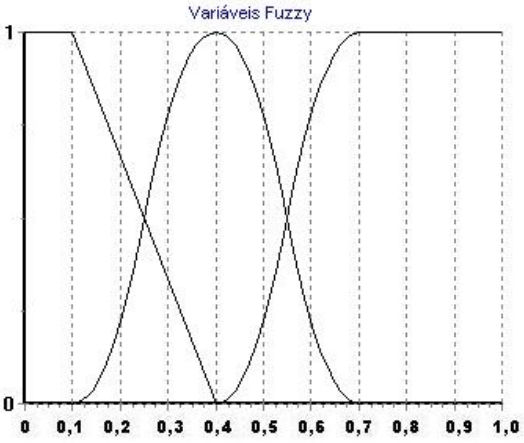


Figura D.6: Combinação 6 (Ver Tabela 4.6).

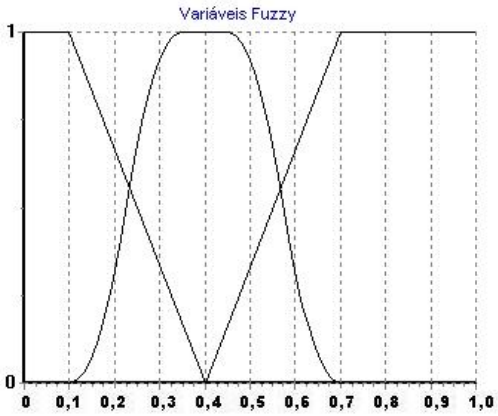


Figura D.7: Combinação 7 (Ver Tabela 4.6).

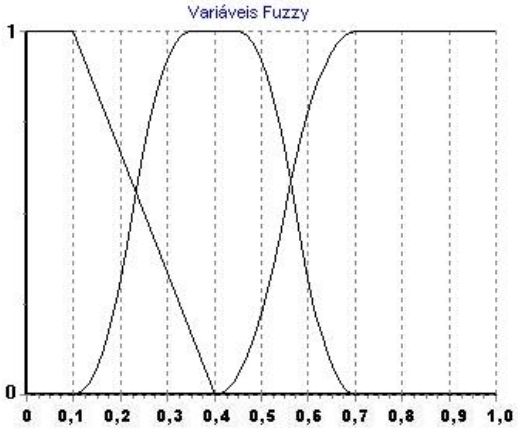


Figura D.8: Combinação 8 (Ver Tabela 4.6).

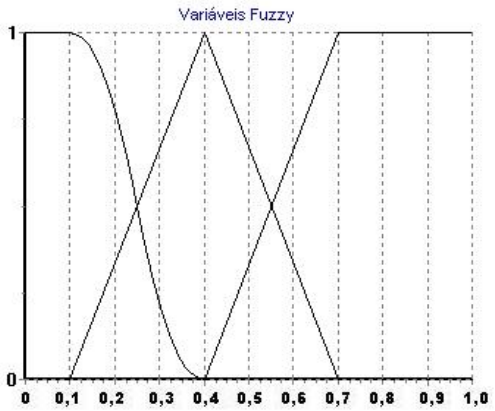


Figura D.9: Combinação 9 (Ver Tabela 4.6).

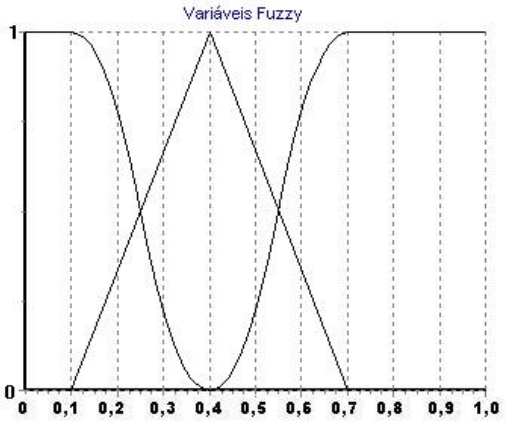


Figura D.10: Combinação 10 (Ver Tabela 4.6).

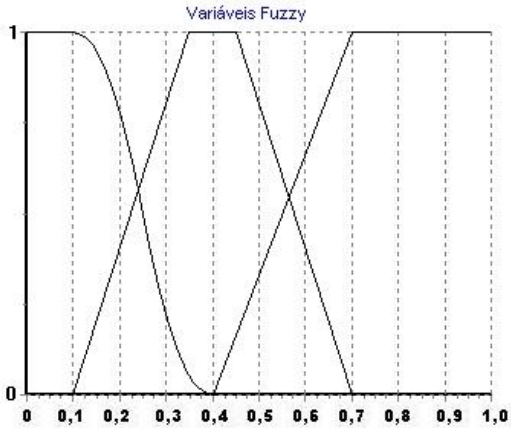


Figura D.11: Combinação 11 (Ver Tabela 4.6).

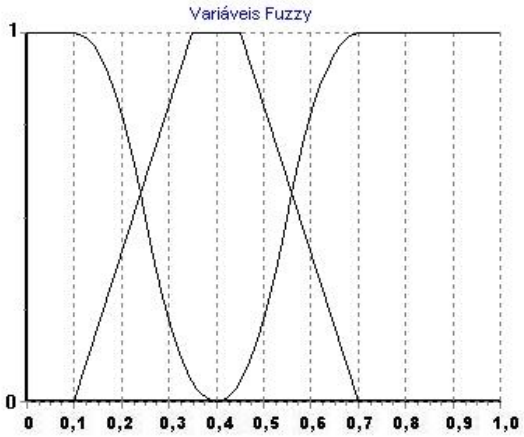


Figura D.12: Combinação 12 (Ver Tabela 4.6).

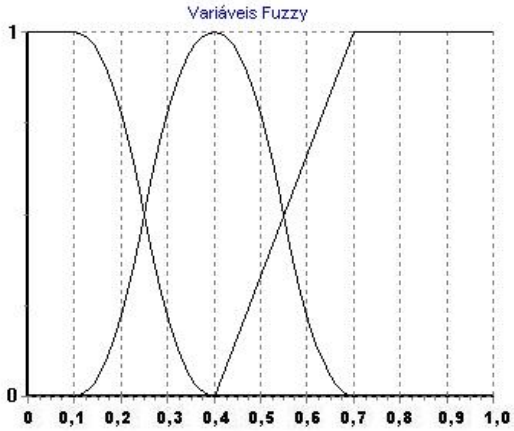


Figura D.13: Combinação 13 (Ver Tabela 4.6).

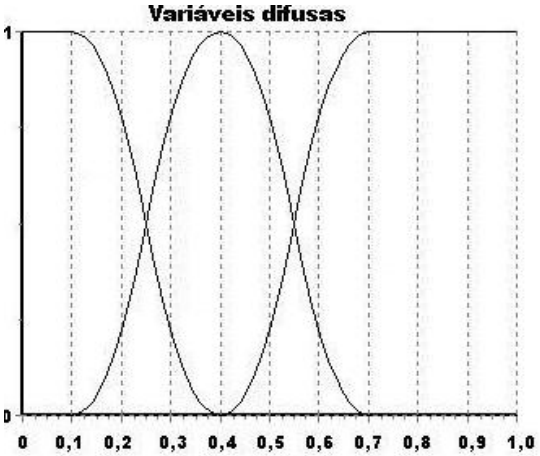


Figura D.14: Combinação 14 (Ver Tabela 4.6).

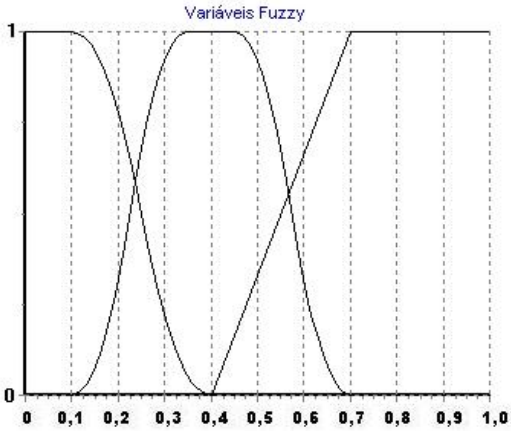


Figura D.15: Combinação 15 (Ver Tabela 4.6).

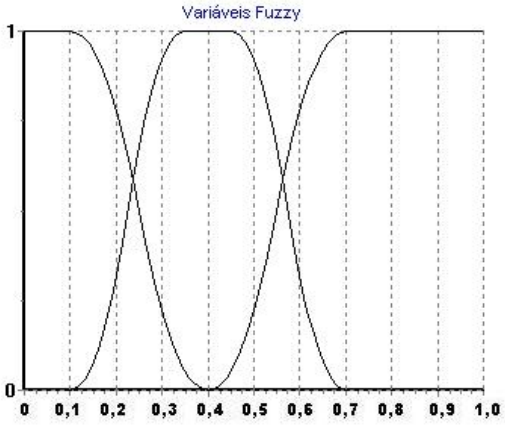


Figura D.16: Combinação 16 (Ver Tabela 4.6).